Universidad Autónoma de Madrid
Escuela politécnica superior

Grado en Ingeniería Informática

Trabajo Fin de Grado

LA EVOLUCIÓN DE DEEPFAKES: ANÁLISIS A NIVEL DE RENDIMIENTO E IMAGEN

Autor: Sergio Romero Tapiador
Tutor: Rubén Tolosana Moranchel
Ponente: Julián Fiérrez Aguilar

JUNIO 2020

LA EVOLUCIÓN DE DEEPFAKES: ANÁLISIS A NIVEL DE RENDIMIENTO E IMAGEN

Autor: Sergio Romero Tapiador
Tutor: Rubén Tolosana Moranchel
Ponente: Julián Fiérrez Aguilar

Biometrics and Data Pattern Analytics - BiDA Lab
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
JUNIO 2020

Resumen

Resumen

En este Trabajo de Fin de Grado se realiza un estudio basado en el análisis a nivel de rendimiento e imagen de vídeos generados por técnicas de manipulación denominadas *DeepFakes* y *Face Swap*. Para tal fin, se han utilizado las bases de datos y técnicas de detección en el estado del arte. Además, se han propuesto nuevos esquemas de detección a nivel de regiones faciales.

En primer lugar, se ha diseñado una arquitectura capaz de procesar imágenes y detectar aquellas que han sido manipuladas. Para ello, se han seleccionado los sistemas más aptos para estas tareas: por un lado, se ha implementado un sistema que procesa imágenes y segmenta las regiones faciales de una cara; por otro lado, se han replicado tres sistemas diferentes encargados de la detección de imágenes *fake*. En este caso, se han escogido los sistemas con mejor rendimiento según el estado del arte, donde han destacado aquellos que utilizan arquitecturas basadas en *Convolutional Neural Networks* (CNN) y las recientes *Capsule Networks*.

Una vez replicados los sistemas de detección de manipulaciones, se ha llevado a cabo su adaptación para las bases de datos analizadas. Estas bases de datos, que contienen vídeos reales y vídeos *fake*, se han dividido en dos generaciones distintas debido a las mejoras de calidad observadas como evolución de las técnicas de manipulación (p.ej. *DeepFakes*).

En tercer lugar, se ha establecido un protocolo experimental para cada sistema y base de datos. Posteriormente, se han entrenado los modelos por regiones faciales y finalmente, se ha elaborado un análisis en base a los rendimiento obtenidos, comparando los diferentes modelos faciales y sistemas evaluados. Estos resultados han generado un artículo de investigación enviado al *IEEE International Conference on Pattern Recognition 2020*.

Por último, se han extraído las conclusiones de este trabajo y a continuación se han propuesto líneas de investigación posibles para futuros trabajos.

Palabras Clave

Fake News, *DeepFakes*, Manipulación Facial, Detección de Manipulaciones, Reconocimiento Facial, Regiones Faciales, Biometría.

Abstract

In this Bachelor's Degree Final Project, we perform an in-depth analysis of digital face manipulation and fake detection techniques in terms of performance and facial regions.

Firstly, it's been designed an architecture capable of processing images and detecting the fake ones. To achieve this goal, the following approaches are developed: a module in order to process images and segment them into different facial regions; a module able to identify both real and fake images. In this case, state-of-the-art systems based on Convolutional Neural Networks (CNN) and Capsule Networks are chosen.

Secondly, the final architecture is implemented and adapted to the databases analysed. These databases, which contain both real and fake videos, are divided into two differents generations due to the quality improvements among them (as an evolution of forgery techniques such as DeepFakes).

Thirdly, an experimental protocol is proposed for each system and database. Then, after training the models, an analysis based on performance and facial regions is elaborated by comparing the different face models, highlighting both the best and worst results. The results obtained in this Bachelor's Degree Final Project has resulted in a research paper sent to the *IEEE International Conference on Pattern Recognition 2020*.

Finally, some conclusions and possible future work lines are drawn.

Key words

Fake News, DeepFakes, Face Manipulation, Fake Detection, Face Recognition, Face Regions, Biometrics

Agradecimientos

En primer lugar, me gustaría agradecer a Rubén Tolosana por toda la confianza y esfuerzo puesto en mí durante todo este año. Por todas las veces que me he equivocado y él ha sabido guiarme por el buen camino. Por todos los obstáculos que me he encontrado y él me ha dicho cómo debería de afrontarlos. Sin duda, este trabajo habría sido totalmente diferente sin su apoyo constante.

En segundo lugar, agradecer a Rubén Vera y a Carlos González por haberme dado la oportunidad de formar parte de BiDA Lab. También, me gustaría dar las gracias a cada uno de los miembros de este grupo que me han estado acompañando en este año tan singular, ayudándome en todo lo posible y sacándome una sonrisa en cualquier momento.

Además, dar las gracias a todas las personas que me han apoyado estos cinco años en la universidad: desde numerosos profesores con ganas de enseñar hasta todos los trabajadores de la cafetería -especialmente a Diego-, sin olvidarme de todos los compañeros que he conocido y que muchos de ellos se han convertido en amistad. Agradecer a todos los profesores que tuve en el colegio e instituto y que me han marcado el camino correcto para llegar hasta aquí.

Por otra parte, agradecer a toda mi familia el apoyo incondicional y constante: a mi madre, por soportarme estos 22 años y por dar todo y más de lo que una madre puede dar; a mi padre, por darme apoyo desde la distancia con su alegría, positividad e ilusión por todo lo que hago; a mi hermana, por quererme y darme fuerzas también desde la distancia. A Irene, por su cariño y apoyo constante y a toda su familia, por darme una segunda casa.

En último lugar -y no por ello menos importante- dar las gracias a todos mis amigos que se han convertido en la segunda familia. A Nacho, Carlos, Sergio y Alberto por todos los viajes, experiencias y por todo lo que venga en el futuro. A toda la familia creada en Toledo: Jorge, Fer, Felipe, Álvaro, Adri, Dani, Víctor, Timothy, Antonio, Kike y Vicente. A toda la familia surgida en Colmenar Viejo y alrededores: Expo, Tomás, Naran, Álvaro, Jaime, Andrés, Alfonso, Blanca, Juan, Jesús, María, Laura, Manuel, Miguel, Alberto, Guille, Mario y Raúl. Y a todos los que no nombro pero que se cruzaron en mi camino y tienen sitio en esta familia tan amplia.

Muchas gracias a todos por formar parte de mi vida.

Sergio Romero
Junio 2020

Índice general

Índice de Figuras

Índice de Tablas

1

Introducción

La llegada incipiente de la era de las *fake news* ha provocado un revuelo entre la sociedad actual [1]. Hoy en día, una noticia falsa puede ser vista por millones de personas mucho antes de que pueda ser identificada y catalogada como tal. Además de artículos ficticios, han aparecido imágenes y vídeos no reales con propósitos fraudulentos. Estos contenidos multimedia se conocen como ***DeepFakes*** y este término ha ido creciendo en popularidad con el paso del tiempo a un ritmo vertiginoso[1].

Cuando se habla de *DeepFakes*, se refiere a las técnicas basadas en *Deep Learning* que permiten crear imágenes y vídeos falsos intercambiando la cara de una persona en un vídeo a partir de la cara de una segunda persona. Este término apareció por primera vez a finales de 2017 en la conocida página web *Reddit*, donde un usuario con ese mismo nombre declaraba haber desarrollado un algoritmo de aprendizaje automático basado en *Deep Learning* capaz de intercambiar las caras de celebridades en vídeos pornográficos [2]. A partir de ese momento, han emergido nuevos escenarios y tipos de ataques donde el centro de atención ha residido en personajes famosos -especialmente actores y políticos-. Esta técnica ha ganado popularidad en los últimos meses debido a la gran calidad y realismo del contenido y así como del uso de herramientas de aprendizaje automático que permiten obtener estos resultados. Otro aspecto importante reside en el uso de sistemas de reconocimiento facial que integran hoy en día muchas aplicaciones: desbloqueo facial, control de acceso o incluso como verificación en métodos de pago. Aunque estos avances permiten numerosas ventajas, son atraídos por agentes maliciosos que manipulan imágenes con el objetivo de sobrepasar cualquier control de seguridad y usurpar la identidad de la persona atacada [3].

Aparte del intercambio de caras entre dos personas, comúnmente llamado ***Identity Swap***, han ido surgiendo nuevos tipos de manipulación facial como se detalla en el Estado del Arte (véase Capítulo 2): generación de caras de personas inexistentes, intercambio de expresiones o incluso síntesis de voz, entre otros. La proliferación de estos ataques ha generado una preocupación a nivel global en el intento de detectar contenido no real. Por ello, las técnicas de detección están siendo cada vez más eficaces ante una imagen o un vídeo falso independientemente de su origen.

Finalmente, estas técnicas han tenido una enorme repercusión e impacto social: por un lado, grandes empresas tecnológicas como Google [4] o Facebook [5] han generado y publicado multitud

[1] https://www.bbc.com/news/technology-49961089

de vídeos *fake*. Junto a Twitter, estas empresas trabajan en la detección de esta nueva tecnología [6]; por otro lado, algunas plataformas de redes sociales permiten a los usuarios crear y compartir este contenido con gran facilidad, como es el caso de *Snapchat*, *TikTok*, *Instagram* o *Zao* [7].

1.1. Motivación del Proyecto

Hoy en día, existen decenas de programas y aplicaciones que permiten generar contenido falso de manera automática sin tener un conocimiento previo del proceso. Es por ello que cualquier persona puede llegar a difundir vídeos *fake* de una persona haciendo o diciendo algo que nunca ha hecho, con consecuencias muy graves para la persona atacada tanto a nivel social como laboral. Estos efectos pueden ser devastadores para la democracia y la seguridad de un país [8].

Al mismo tiempo, la Inteligencia Artificial (*IA*) juega un papel fundamental y en concreto, algoritmos de aprendizaje automático (*Machine Learning*) basados en técnicas de *Deep Learning*. Son las **redes neuronales profundas** -junto con otros métodos- las que generan este contenido multimedia que engaña tanto al ojo humano como a las máquinas más avanzadas. Además, son cada vez más precisas, dificultando la labor de la detección y mejorando la calidad del contenido visual. No obstante, estos sistemas también son utilizados para la detección de *DeepFakes*, obteniendo rendimientos muy superiores al que una persona experimentada puede alcanzar [4].

Por estos motivos este proyecto se enmarca en el campo de la **detección de técnicas digitales de manipulación facial, denominadas *DeepFakes***. Este trabajo supone un primer estudio hacia la detección de dichas técnicas, observando su creciente evolución y aparición de nuevas bases de datos, así como el análisis a nivel de rendimiento y calidad de la imagen. Los resultados obtenidos en este Trabajo Final de Grado han generado un artículo de investigación enviado al *IEEE International Conference on Pattern Recognition 2020*, y se encuentra disponible en *ArXiv* [9].

1.2. Objetivos y Enfoque

En la actualidad hay disponibles multitud de bases de datos de acceso libre que, junto con la evolución de las técnicas de *Deep Learning*, permiten la generación de contenidos falsos de gran calidad con sus correspondientes implicaciones sociales y legales que eso puede llevar.

Por ello, el desarrollo de este trabajo se centra los siguientes objetivos:

- Estudio de las técnicas digitales de manipulación facial y de detección de las mismas que conforman el estado del arte, en concreto, las basadas en *Identity Swap*.
- Análisis de la evolución de *DeepFakes*: estudio en profundidad de las bases de datos pertenecientes a esta técnica, así como de los sistemas de detección.
- Análisis de nuevos sistemas de detección de contenido falso mediante técnicas basadas en *Convolutional Neural Networks (CNN)* y *Capsule Networks* considerando las distintas regiones faciales.
- Evaluación exhaustiva de los sistemas desarrollados, obteniendo un análisis por técnica y base de datos.
- Estudio de metodologías y sistemas que posibiliten una mejora del trabajo propuesto en un futuro.

1.3. Metodología y Plan de Trabajo

Para cumplir satisfactoriamente con los objetivos propuestos en la anterior sección, se ha seguido un plan de trabajo de manera estricta:

- **Estudio del estado del arte:** antes de empezar un proyecto, el primer paso radica en adquirir la formación necesaria para aplicar los conocimientos una vez se han estudiado. Para este proyecto, se ha llevado a cabo un profundo estudio de las diferentes técnicas de manipulación y detección de *Identity Swap*. Además, se han examinado las diferentes bases de datos que lo conforman.
- **Estudio y desarrollo del software existente:** para poder aplicar los conocimientos obtenidos a lo largo del estudio realizado, se ha de haber analizado las diferentes arquitecturas y algoritmos necesarios para el desarrollo de los protocolos y del código que se vaya a utilizar. Particularmente, parte del código se encuentra públicamente en la plataforma *GitHub*. Tras el estudio, se han implementado: algoritmos para la segmentación de caras y de sus distintas regiones faciales, sistemas de detección a partir de CNN y *Capsule Networks*, protocolos experimentales y todo el código necesario para entrenar y evaluar los sistemas. Se han utilizado las librerías de *Tensorflow*, *Keras* y *PyTorch*, la herramienta *OpenFace* [10] para el reconocimiento facial y el código se ha implementado en *Python* y *MatLab*.
- **Evaluación de los experimentos y resultados obtenidos:** se ha establecido un protocolo uniforme para cada sistema y base de datos. Posteriormente, una vez se han ejecutado todos los procesos anteriores y obtenido los resultados, el siguiente paso a realizar consiste en analizar los distintos rendimientos que se han logrado para cada una de las bases de datos, comparando las diferencias que ofrecen los distintos sistemas. Paralelamente, esta comparación se lleva a cabo con el estado del arte.
- **Publicación de artículos de investigación:** los resultados obtenidos en este Trabajo de Fin de Grado han generado un artículo de investigación enviado al *IEEE International Conference on Pattern Recognition 2020*, y se encuentra disponible en *ArXiv* [9].
- **Redacción de la memoria:** en último lugar, tras realizar un exhaustivo estudio y posteriormente proceder al desarrollo y evaluación del trabajo en cuestión, se elabora la presente memoria, detallando cada uno de los puntos expuestos.

1.4. Organización de la Memoria

El presente trabajo se estructura de la siguiente manera:

- **Capítulo 1:** Introducción.
- **Capítulo 2:** Estado del arte.
- **Capítulo 3:** Sistemas propuestos.
- **Capítulo 4:** Desarrollo experimental.
- **Capítulo 5:** Conclusiones finales y trabajo futuro.

2
Estado del arte

En este capítulo se muestran los diferentes tipos de manipulación digital que existen en la actualidad, entre los que destacan ***Face Synthesis***, ***Attribute Manipulation***, ***Expression Swap*** e ***Identity Swap***. Tras el análisis inicial de los distintos tipos de manipulación facial, se describe en detalle el estado del arte de *Identity Swap*, que constituye el principal objetivo de estudio del presente Trabajo de Fin de Grado.

2.1. Tipos de Manipulación: Inicios y Evolución

En los últimos años, las técnicas que generan y manipulan contenido multimedia han ido progresando notablemente, obteniendo resultados tan reales que apenas pueden ser distinguibles por el ser humano, como es el ejemplo del rostro de una persona[1]. Además, todo ello ha crecido enormemente debido a dos principales razones: *i)* el acceso a grandes conjuntos de datos, y *ii)* la evolución de técnicas de *Deep Learning* [11] que eliminan procesos manuales de edición. Estas herramientas, como ***Autoencoders*** (AE) o las redes **Generative Adversarial Nets** (*GAN*) [12] [13], han propiciado la creación de este contenido *fake*, desde imágenes de personas no reales hasta intercambio de caras en vídeos[2]. A continuación se describen algunas de las técnicas de manipulación facial más populares en el estado del arte.

2.1.1. *Face Synthesis*

Esta manipulación consiste en crear imágenes de caras de personas no existentes mediante redes GAN, consiguiendo resultados que demuestran un alto realismo[3], como se puede observar en la Figura 2.1 (izquierda). Las redes GAN han ido evolucionando dando lugar a arquitecturas más especificas para la síntesis de caras, como es el caso de *StyleGAN* . Por otro lado, a pesar del realismo que presentan estas imágenes, cada una de ellas está caracterizada por huellas específicas producidas por las redes GAN, lo que ayuda en las labores de detección [14]. No obstante, la red *GAN fingerprint Removal* (*GANprintR*) provee una arquitectura capaz de eliminar estas huellas [15].

[1] http://www.whichfaceisreal.com/
[2] https://www.youtube.com/channel/UCKpH0CKltc73e4wh0$_pgL3g$
[3] https://thispersondoesnotexist.com/

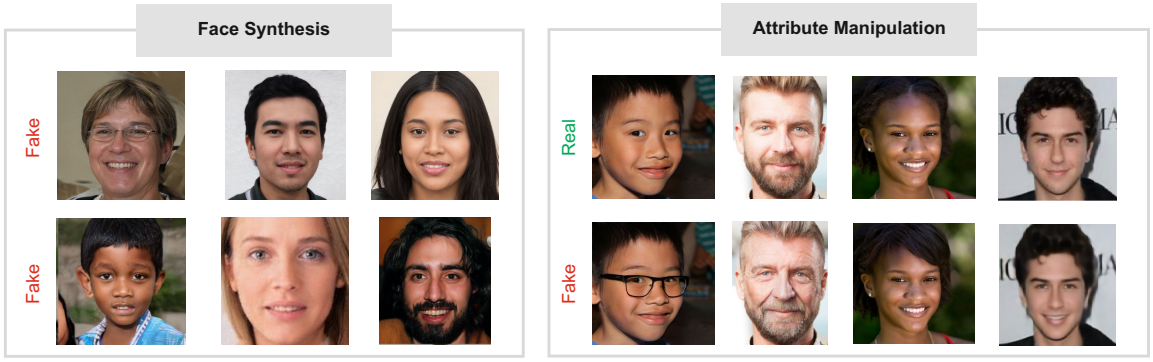


Figura 2.1: Ejemplos de imágenes reales y *fake* para el tipo de manipulación facial *Face Synthesis* (izquierda) y *Attribute Manipulation* (derecha). Imágenes extraídas de [1, 3, 4] y [14].

Entre los diferentes sectores en los que se puede aplicar esta técnica se encuentran las industrias de videojuegos y de modelado 3D. Por otro parte, también puede llegar a ser un peligro en el ámbito social, como es la creación de perfiles falsos con apariencia real en redes sociales con el objetivo de generar desinformación [2].

2.1.2. *Attribute Manipulation*

Esta manipulación consiste en modificar atributos de la cara: desde el color del cabello o de la piel, cambio de género o cambio de edad hasta la posibilidad de añadir accesorios como gafas (véase Figura 2.1 -derecha-). Al igual que *Face Synthesis*, este tipo de manipulación se lleva a cabo mediante redes *GAN* y novedosas variantes: *attGAN* permite elegir qué atributos se pueden añadir o eliminar [16]; *StarGAN* y *STGAN*, por otro lado, ofrecen una mejora visual respecto a anteriores arquitecturas [17] [18].

Hoy en día, este tipo de manipulación se ha integrado en algunas industrias, donde muchos de los productos que venden -como cosméticos, gafas o diferentes peinados- pueden llegar a ser probados en un entorno virtual. Una de las aplicaciones más famosas a nivel mundial es *FaceApp*, donde permite realizar algunas de estas modificaciones[4].

2.1.3. *Expression Swap*

Esta manipulación consiste en modificar la expresión facial de una persona, reemplazándola por la expresión del rostro de otra. Entre las técnicas más populares destacan *Face2Face* [19] y *NeuralTextures* [20], y es importante remarcar que este tipo de manipulación puede tener un impacto social y político enorme: esta técnica genera vídeos en los que una persona dice algo que realmente no ha pronunciado nunca[5] -normalmente los ataques van dirigidos a celebridades como políticos y actores-.

Actualmente, únicamente *FaceForensics++* [4] ha proporcionado una base de datos con vídeos de este tipo de manipulación, como se analiza en la Figura 2.2. En concreto, ha publicado dos diferentes modelos de expresión facial:

- ***Face2Face*** [19]: mediante técnicas de *computer graphics* transfieren la expresión facial de una persona -vídeo original- al rostro de otra persona -vídeo destino-, manteniendo la identidad de la primera persona.

[4]https://www.faceapp.com/
[5]https://www.youtube.com/watch?v=cnUd0TpuoXI

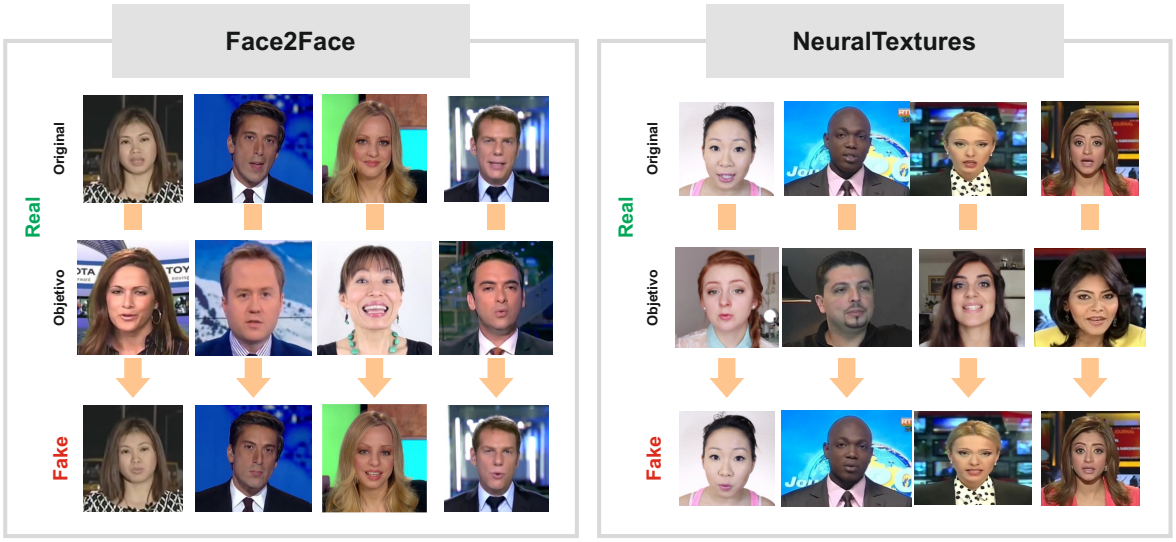


Figura 2.2: Ejemplos de imágenes manipuladas del tipo *Expression Swap*, usando los algoritmos *Face2Face* (izquierda) y *NeuralTextures* (derecha). Las imágenes de los dos tipos de manipulación proceden de la base de datos *FaceForensics++* [4].

- ***NeuralTextures*** [20]: en esta ocasión solo se transfiere la región de la boca, desde un vídeo origen a un vídeo destino.

Aparte de estos 3 tipos de manipulación facial descritos, existen otras técnicas que han ido ganando popularidad con el paso del tiempo [2]:

- ***Identity Swap***: este método consiste en sustituir la cara de una persona por el rostro de otra en un vídeo, tal y como se muestra en la Figura 2.3. Este tipo de manipulación se analiza en la siguiente sección (véase Sección 2.2), **siendo objeto de estudio de este Trabajo de Fin de Grado**.
- ***Face Morphing***: consiste en generar una cara biométrica artificial a partir de los rostros de dos o más personas. Por lo tanto, corresponde con un tipo de manipulación facial a nivel de imagen [21].
- ***Face De-Identification***: esta técnica de manipulación facial se encarga de eliminar la información presente en el rostro de una persona, con el fin de preservar la identidad y privacidad de la misma, tanto a nivel de imagen como de vídeo[6] [22].
- **Sincronización labial**: este método de *Expression Swap* se aplica para la sincronización de labios en un vídeo a partir de un audio o texto de entrada cualquiera [23]. Una vez el sistema ha analizado la forma de la boca (y labios) de una persona al reproducir un sonido, posteriormente se sintetiza para generar un vídeo sincronizando y emparejando un audio o texto con los movimientos labiales de la persona, obteniendo resultados bastante realistas, tal como se muestra en el vídeo manipulado de Barack Obama[7].

[6] https://www.youtube.com/watch?v=cCYnBtni7Wg
[7] https://www.youtube.com/watch?v=cQ54GDm1eL0

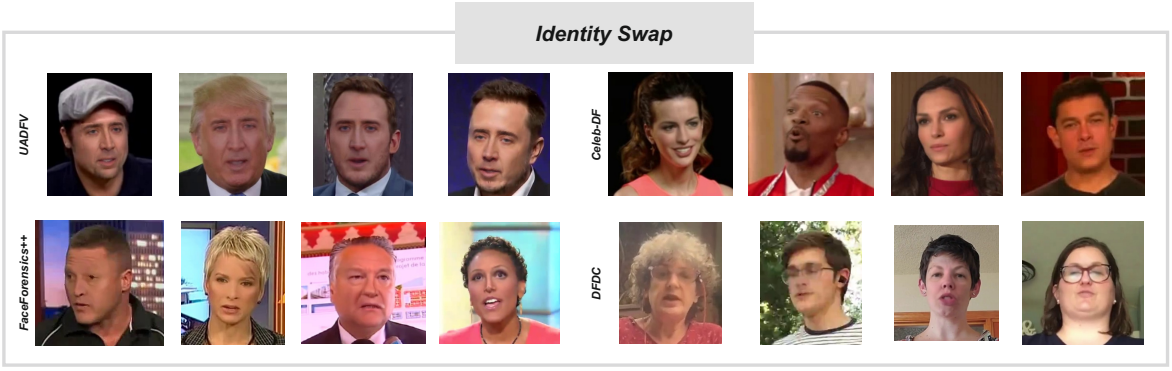


Figura 2.3: Ejemplos de imágenes manipuladas del tipo *Identity Swap*. Las imágenes se han extraído de las bases de datos *UADFV*, *FaceForensics++*, *Celeb-DF* y *DFDC* [4] [5] [27] [29].

2.2. Identity Swap

Esta manipulación consiste en el **reemplazo de la cara de una persona con el rostro de otra**, con el fin de generar vídeos *fake* muy realistas (véase Figura 2.3). Sin duda, es el método de manipulación facial más popular hoy en día, tanto en el ámbito de investigación como en el ámbito social. Los numerosos estudios y las decenas de aplicaciones al alcance de cualquiera ha generado una evolución notable en las técnicas de manipulación y detección. Por estos motivos, **este trabajo se ha centrado en el análisis y desarrollo de este tipo de contenido falso**.

2.2.1. Técnicas de Manipulación

Aunque existan diversas aplicaciones móviles y programas a nivel de usuario (como es el caso de la *app* china ZAO), los verdaderos progresos se han encontrado en las técnicas de manipulación que han dado lugar a diferentes bases de datos públicas.

Técnicas Basadas en redes GAN

Esta primera técnica de manipulación, empleada para la base de datos DeepfakeTIMIT [24], utiliza un algoritmo de *Face Swap* basado en redes *GAN*[8].

Estas redes denominadas ***Generative Adversarial Networks***, son empleadas también por otros tipos de manipulación como *Face Synthesis* o *Attribute Manipulation*. La arquitectura está formada por dos modelos: un modelo generativo y un modelo discriminativo. El modelo generativo, por un lado, se encarga de generar imágenes lo más reales posibles. El modelo discriminativo, por otro lado, se encarga de distinguir entre imágenes reales e imágenes *fake*, estas últimas creadas por el modelo generativo. Por lo tanto, este proceso se convierte en un juego de dos jugadores *minimax*, donde el modelo generativo intentará crear imágenes cada vez más reales para que el modelo discriminativo no sea capaz de diferenciarlas con las verdaderamente auténticas [13].

Para esta técnica de manipulación, se han usado unas redes generativas denominadas **CycleGAN** [25] y empleado un método ***Multi-Task Cascaded Convolution Networks***(MTCNN) para proporcionar mas estabilidad en las detecciones y alineamientos faciales [26].

[8]https://github.com/shaoanlu/faceswap-GAN

	Base de datos	Vídeos reales	Vídeos *fake*
1ª Generación	*UADFV* (2018) [27]	49 (YouTube)	49 (*FakeApp*)
	DeepFakeTIMIT (2018) [24]	-	620 (*faceswap-GAN*)
	FaceForensics++ (2019) [4]	1.000 (YouTube)	1.000 (*FaceSwap*) 1.000 (*DeepFake*)
2ª Generación	*DeepFakeDetection* (2019) [28]	363 (Actores)	3.068 (*DeepFake*)
	Celeb-DF (2019) [29]	890 (YouTube)	5.639 (*DeepFake*)
	DFDC (2019) [5]	1.131 (Actores)	4.119 (Desconocido)

Tabla 2.1: Bases de datos públicas disponibles para *Identity Swap*. Fuente: [2]

Técnicas Basadas en *Autoencoders*

Los ***Autoencoders*** son un tipo de red neuronal *feedforward* cuyo objetivo reside en obtener una salida lo más idéntica posible a la entrada con la que fue alimentada. De este modo, los datos de entrada se comprimen a un espacio dimensional menor a través del ***encoder***. Una vez obtenido el ***latent feature vector*** o código generado tras pasar por el *encoder*, el ***decoder*** se encargará entonces de reconstruir los datos de entrada lo más semejante posible a partir de este código. Los *Autoencoders* son usados comúnmente para la generación de vídeos *fake* y en concreto, para la técnica *DeepFakes*.

Un ejemplo de ello es la base de datos *FaceForensics++*, la cual emplea estas redes para reconstruir imágenes de caras correspondientes al entrenamiento. Además, su sistema desarrollado utiliza un detector facial que recorta y alinea las imágenes obtenidas, donde estas son posteriormente enviadas a dos *Autoencoders* diferentes: un *Autoencoder* entrenado para las caras originales y otro entrenado para las caras que conformarán las máscaras. Estos dos *Autoencoders* finalmente se adaptan entre sí para generar la imagen final [4].

Por otra parte, los vídeos *fake* obtenidos por la base de datos *Celeb-DF* también han sido generados mediante un algoritmo basado en *DeepFakes*: a partir de un vídeo de entrada, se obtiene la cara en cada secuencia y son enviados al *Autoencoder* -formado por dos (**CNN**)-, el cual se encarga de sintetizar las caras con el fin de emparejar las mismas expresiones faciales del rostro original y el rostro perteneciente a la máscara. Es decir, cuantos más vídeos existan de una única identidad, más expresiones faciales se obtendrán, lo que contribuirá a generar vídeos más realistas a la hora de obtener las caras. Además del ***Autoencoder*** principal, existe un sistema par codificador-decodificador encargado de trabajar en paralelo con el sistema principal para minimizar errores en la reconstrucción de la máscara.

Computer Graphics

Aparte de estas técnicas más comunes, *FaceForensics++* introdujo una nueva basada en *Computer Graphics* denominada *FaceSwap* [4]. Para esta técnica consideraron un algoritmo que permite intercambiar la cara de una persona en el rostro de otra, gracias en parte a la utilización de técnicas de alineamiento facial, algoritmos fundamentados en el método de *Gauss-Newton* y técnicas basadas en *image blending*[9].

[9]https://github.com/MarekKowalski/FaceSwap

Figura 2.4: Imperfecciones y debilidades observadas en la 1ª generación de *Identity Swap*. Las imágenes se han extraído de las bases de datos *UADFV* y *FaceForensics++* [4] [27].

2.2.2. Bases de Datos Públicas

Desde las primeras bases de datos como ***UADFV*** [27] o ***DeepFakeTIMIT*** [24] hasta las más recientes como ***Celeb-DF*** [29] o ***DFDC*** [5], se han llevado a cabo numerosas mejoras visuales, incrementando el realismo de los vídeos *fake*. Debido a estas grandes diferencias entre bases de datos, se han dividido en dos generaciones distintas atendiendo a las características y técnicas usadas en cada una de ellas. Mientras que en la primera generación se detecta con bastante facilidad los vídeos que han sido manipulados, para la segunda generación tanto a nivel humano como a nivel de máquina la tarea se vuelve enormemente compleja, tal y como analizamos en los siguientes capítulos de este Trabajo de Fin de Grado.

En la Tabla 2.1 se observan las diferentes bases de datos que conforman *Identity Swap* en la actualidad, agrupadas por las dos generaciones mencionadas. En general, los vídeos *fake* de la 1ª generación se caracterizan por: *i)* baja calidad en las caras sintetizadas, *ii)* alto contraste de color entre la máscara *fake* sintetizada y la piel de la cara original, *iii)* contorno visible en la máscara *fake*, *iv)* elementos faciales visibles del vídeo original, *v)* pocas variaciones en la pose durante el vídeo, y *vi)* aparición de extraños artefactos entre las secuencias consecutivas de un vídeo, como se examina en la Figura 2.4. Además, estas primeras bases de datos públicas solo consideran escenarios controlados en cuanto a la posición de la cámara (suele estar fija) y las condiciones de luz (buena iluminación y pocos cambios de la intensidad de la misma). Muchos de estos aspectos mencionados han sido mejorados en las posteriores bases de datos, las que forman parte de la segunda generación. En efecto, bases de datos como ***DFDC*** [5] ofrecen diferentes escenarios y situaciones: desde vídeos grabados en el exterior en condiciones lumínicas muy bajas (como puede ser de noche), hasta situaciones en las que la persona se mueve físicamente durante el vídeo, con variaciones en la pose y en ocasiones ocultando parte de su cara [9]. En la Figura 2.5 se observan las mejoras y diferentes condiciones de captura consideradas en la 2ª generación.

1ª Generación

La primera base de datos disponible fue ***UADFV*** [27], formada por 49 vídeos reales procedentes de YouTube -en estos vídeos figuran personajes famosos, políticos y actores, todos ellos **varones**-, los cuales fueron usados para generar los 49 vídeos *fake* en donde se intercambian la

Figura 2.5: Mejoras y aparición de nuevos escenarios de la 2ª generación respecto de la 1ª. Las imágenes se han extraído de las bases de datos *Celeb-DF* y *DFDC* [5] [29].

cara únicamente con un actor, en este caso Nicolas Cage. Para esta primera base de datos, se utilizó el programa ***FakeApp***[10]. También en 2018 se publicó la base de datos ***DeepFakeTIMIT*** [24], formada por 620 vídeos *fake* de 32 sujetos diferentes. Este contenido fue generado a partir de los vídeos originales procedentes de de la base de datos *VidTIMIT* [30] mediante las técnicas basadas en redes GAN comentadas previamente.

La última base de datos que pertenece a la 1ª generación es una de las más populares, denominada ***FaceForensics++***[4]. Junto con *UADFV*, se detallan algunas imágenes que conforman estas bases de datos en la Figura 2.6. En primer lugar, escogieron 1000 vídeos de YouTube -más de la mitad corresponden con vídeos protagonizados por mujeres, relacionadas con el mundo del periodismo- para posteriormente intercambiar las caras de las personas entre ellas. Es decir, para formar los vídeos *fake* tomaron pares de vídeos y cambiaron el rostro de los periodistas entre sí. Como se ha detallado anteriormente, esta base de datos ofrece dos técnicas diferentes: *FaceSwap* y *DeepFakes*, obteniendo 1000 vídeos *fake* por cada técnica desarrollada.

2ª Generación

Incluída en el trabajo realizado por *FaceForensics++*, se publicó la base de datos ***DeepFakeDetection*** [28], la cual corresponde a la segunda generación. Esta base de datos, que contó con el apoyo de Google, consta de 363 vídeos reales de 28 actores pagados en 16 escenarios diferentes. Los 3068 vídeos *fake* fueron generados por un algoritmo de *Identity Swap* basado en la técnica *DeepFake* con una mejora notable frente a las bases de datos la anterior generación.

No obstante, *Celeb-DF* y *DFDC* corresponden con las bases de datos más realistas hasta el momento, ambas publicadas en 2019 (véase Figura 2.7). Por un lado, ***Celeb-DF*** [29] contiene 890 vídeos reales (obtenidos de YouTube) y un total de 5639 *fake* de diferentes actores conocidos. Específicamente, corresponden con entrevistas a 59 celebridades de diferentes grupos étnicos y géneros: el **56,8 %** pertenece a hombres y el **43,2 %** a mujeres; el 88,1 % pertenecen a una etnia caucásica, el 6,8 % a una etnia afroamericana y el resto (5,1 %), corresponden a una etnia asiática. De este modo, se han conseguido mejoras en: *i)* caras sintetizadas en baja resolución, *ii)*

[10]https://www.malavida.com/en/soft/fakeapp/

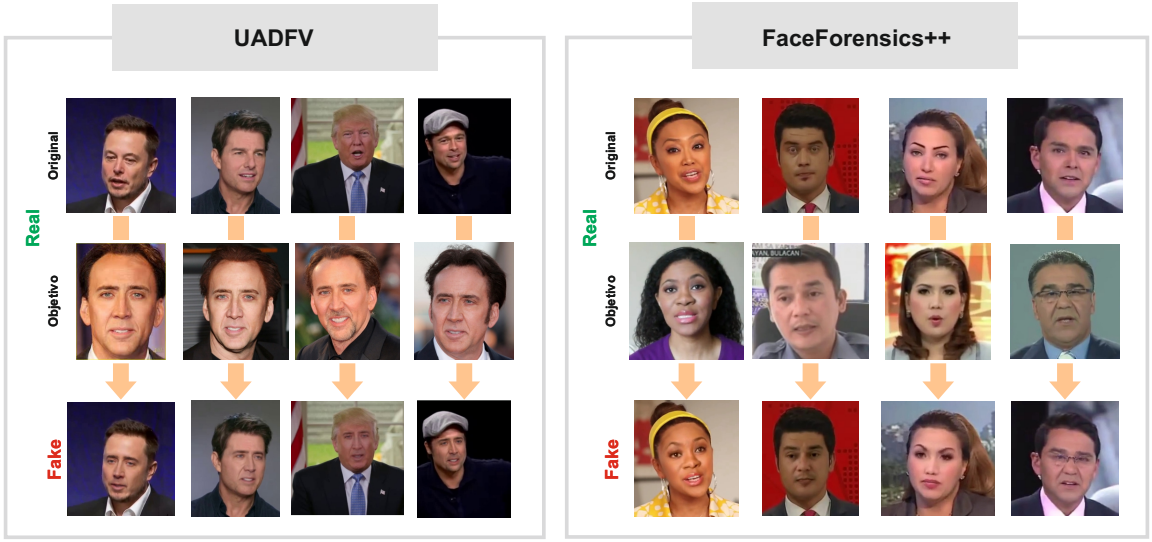


Figura 2.6: Ejemplo de imágenes *fake* correspondientes a la 1ª generación. Las imágenes -tanto reales como *fake*- se han extraído de las bases de datos *UADFV* y *FaceForensics++*, respectivamente [4] [27].

inconsistencia de color entre la máscara y la cara original, *iii)* partes visibles de la cara original, y *iv)* parpadeo temporal entre *frames* -mejoras de la máscara para obtener resultados consistentes entre secuencias consecutivas- [29].

En último lugar, se encuentra la base de datos ***DFDC*** o ***DeepFake Detection Challenge*** [5], publicada por Facebook en colaboración con otras empresas como Microsfot, Amazon o el Instituo Tecnológico de Massachusetts (MIT). Aparte de publicar una base de datos, lanzaron un reto global en el que se animaba a crear el mejor sistema de detección a cambio de una recompensa económica[11]. Esta base de datos contiene 1131 vídeos reales de 66 actores pagados, generando un total de 4119 vídeos *fake* -no especifican los algoritmos desarrollados para la técnica de manipulación-. En este caso, el **74 %** de los actores corresponden con mujeres y el **26 %**, con hombres. Además, el 68 % de ellos pertenecen a una etnia caucásica, el 20 % a una etnia afroamericana, el 9 % a una etnia del este asiático y el 3 %, corresponde a una etnia del sur asiático. Como se ha comentado previamente, los vídeos se han grabado en **diversos escenarios**, distintas condiciones de luz (de día y de noche), diferentes distancias entre la cámara y la persona o variaciones en la pose, lo que dificulta el proceso de detección.

Cabe destacar que algunas bases de datos como *FaceForensics++* han modificado la calidad de los vídeos falsos, proporcionando hasta 3 calidades diferentes -calidad original (*Raw*), alta calidad (HQ) y baja calidad (LQ)- demostrando que la calidad de los mismos afecta tanto a la manipulación como a la detección, además de que visualmente se aprecia considerablemente.

2.2.3. Técnicas de Detección

En esta sección se analizan las diferentes técnicas que se han utilizado para la detección de vídeos falsos de tipo *Identity Swap*, detallando los métodos más comunes y con mejores rendimientos actuales, tal y como se muestra en la Tabla 2.2. Por un lado, los mejores resultados para cada base de datos se han marcado en **negrita** y, por otro lado, se han resaltado en *itálica* aquellos resultados obtenidos con sistemas de detección entrenados con otras bases de datos

[11] https://deepfakedetectionchallenge.ai/

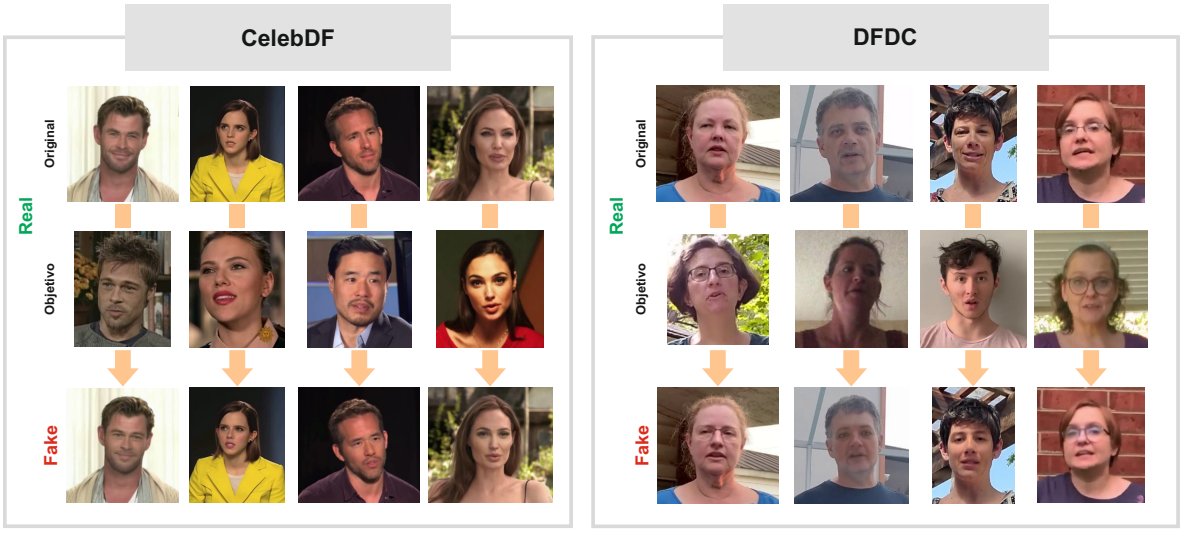


Figura 2.7: Ejemplo de imágenes *fake* correspondientes a la 2ª generación.. Las imágenes -tanto reales como *fake*- se han extraído de las bases de datos *Celeb-DF* y *DFDC*, respectivamente [5] [29].

distintas a las evaluadas, con el objetivo de medir su capacidad de generalización a nuevos tipos de ataques. Cabe destacar que se han utilizado diferentes métricas para cada estudio, por lo que complica enormemente la comparación entre los diferentes sistemas, además de que cada uno ha llevado a cabo un protocolo distinto al resto:

- **Acc** (*accuracy* o exactitud): esta métrica nos indica el porcentaje de aciertos que ha tenido el sistema -imágenes reales identificadas como reales e imágenes *fake* identificadas como *fake*- frente al número total de imágenes evaluadas (tanto reales como falsas).
- **AUC** (*Area Under the Curve* o Área Bajo la Curva): se define como la medida del rendimiento frente a problemas de clasificación para distintos umbrales. Esta métrica se representa mediante la curva *Receiver Operating Characteristic* (ROC), mostrando la capacidad que tiene el modelo de diferenciar entre clases. La curva ROC proyecta el ratio de verdaderos positivos (VPR) frente al ratio de falsos positivos (FPR) [31].
- **EER** (*Equal Error Rate* o tasa de error igual): esta métrica, que normalmente es utilizada en seguridad biométrica, nos determina la tasa de error del sistema cuando la Falsa Aceptación (FA) y el Falso Rechazo (FR) son iguales. Por un lado, la Falsa Aceptación corresponde cuando un sistema identifica erróneamente que una imagen *fake* es real. Por otro lado, el Falso Rechazo ocurre cuando un sistema identifica de manera errónea que una imagen real es *fake* (opuestamente a FA). Finalmente, esta tasa se suele representar a través de las curvas DET (*Detection Error Trade-off*) [32].

Medidas de Calidad de la Imagen

Korshunov y Marcel [30] mostraron diferentes sistemas de detección en su estudio utilizados para la base de datos **DeepFakeTIMIT**. Consideraron un primer sistema de detección audio-visual a partir de inconsistencias entre los movimientos de labios y el audio en cuestión, con el propósito de distinguir entre un vídeo real -el movimiento de labios y audio estarían sincronizados- y uno manipulado -en este caso podrían no estarlo-, pero tras evaluar la base de

Tabla 2.2: Estado del Arte de los diferentes métodos de detección en *Identity Swap*. Los mejores resultados obtenidos para cada base de datos están remarcados en **negrita**. Por otro lado, se señalan en *itálica* aquellas bases de datos que no han sido entrenadas para el método al que pertenece, con el objetivo de medir la capacidad de generalización del estudio. FF++ = *FaceForensics++*, AUC = *Area Under the Curve*, ACC. = *Accuracy*, EER = *Equal Error Rate*. Fuente: [2] [29].

Método	**Clasificador**	**Mejor Resultado**	**Base de Datos**
Medidas de Calidad de la Imagen [24]	SVM	**EER = 3,3 %**	**DeepFakeTIMIT (LQ)**
		EER = 8,9 %	**DeepFakeTIMIT (HQ)**
Artefactos Visuales [33]	Regresión Logística MLP	*AUC = 70,2 %*	*UADFV*
		AUC = 77,0 %	*DeepfakeTIMIT (LQ)*
		AUC = 77,3 %	*DeepfakeTIMIT (HQ)*
		AUC = 78,0 %	*FF++ / DFD*
		AUC = 66,2 %	*DFDC Preview*
		AUC = 55,1 %	*Celeb-DF*
Pose y Expresiones Faciales [34]	SVM	AUC = 89,0 %	UADFV
		AUC = 55.1 %	*DeepfakeTIMIT (LQ)*
		AUC = 53,2 %	*DeepfakeTIMIT (HQ)*
		AUC = 47,3 %	*FF++ / DFD*
		AUC = 55,9 %	*DFDC Preview*
		AUC = 54,6 %	*Celeb-DF*
DSP-FWA [29] [35]	CNN	**AUC = 97,7 %**	**UADFV**
		AUC = 99,9 %	**DeepfakeTIMIT (LQ)**
		AUC = 99,7 %	**DeepfakeTIMIT (HQ)**
		AUC = 93,0 %	*FF++ / DFD*
		AUC = 75,5 %	***DFDC Preview***
		AUC = 64,6 %	***Celeb-DF***
XceptionNet (Rössler *et al.*) [4]	CNN	Acc. ≃ 94,0 %	FF++ (DeepFake, LQ)
		Acc. ≃ 98,0 %	**FF++ (DeepFake, HQ)**
		Acc. ≃ 100,0 %	**FF++ (DeepFake, RAW)**
		Acc. ≃ 93,0 %	FF++ (FaceSwap, LQ)
		Acc. ≃ 97,0 %	**FF++ (FaceSwap, HQ)**
		Acc. ≃ 99,0 %	**FF++ (FaceSwap, RAW)**
XceptionNet (Dolhansky *et al.*) [5]	CNN	**Precision = 93,0 %** **Recall = 8,4 %**	**DFDC Preview**
Capsule Forensic [36]	Capsule Networks	*AUC = 61,3 %*	*UADFV*
		AUC = 78,4 %	*DeepfakeTIMIT (LQ)*
		AUC = 74,4 %	*DeepfakeTIMIT (HQ)*
		AUC = 96,6 %	***FF++ / DFD***
		AUC = 53,3 %	*DFDC Preview*
		AUC = 57,5 %	*Celeb-DF*
DenseNet + Alignment + BiDir [37]	CNN + RNN	**AUC = 96,9 %**	**FF++ (DeepFake, LQ)**
		AUC = 96,3 %	**FF++ (FaceSwap, LQ)**

datos, el resultado fue de 41,8 % de EER. Además, se desarrollaron otras técnicas basados en la imagen, como es el caso de *Image Quality Measures*) (IQM) [38]. No obstante, este método obtuvo resultados de EER de **3,3 %** y **8,9 %** para vídeos de baja y alta calidad, respectivamente. En este sistema vectorial, se usaron 129 diferentes medidas respecto a la calidad de la imagen -relación señal/ruido, especularidad, borrosidad, etc.- Junto a este sistema, se utilizó un clasificador *Support Vector Machine* (SVM), proporcionando mejores resultados para las imágenes con baja calidad (*LQ*).

Artefactos Visuales

Matern *et al.* [33] propusieron un sistema de detección basado en aspectos visuales como detalles que faltan en el área de la boca y en los ojos o el color de estos últimos, entre otros. Para ello, se han desarrollado dos clasificadores diferentes: un modelo de regresión logística y un *Multilayer Perceptron* (MLP). Este sistema fue entrenado y evaluado con una base de datos privada, aunque posteriormente se evaluó la capacidad de generalización frente a las demás bases de datos, obteniendo un valor AUC del **78 %** para *FaceForensics++*.

Pose y Expresiones Faciales

Este sistema basado en expresiones faciales y movimientos de la cabeza propone la detección de vídeos *fake* a partir de la estimación de la pose de la cabeza. En otras palabras, en algunas técnicas de manipulación unen directamente la máscara sintetizada con la cara original sin tener en cuenta la pose de la cabeza, generando una inconsistencia en la misma e introduciendo errores que mediante estimaciones 3D de la pose son detectados. Este sistema desarrollado por Yang *et al.* [27] utiliza *landmarks* (puntos de referencia) faciales para extraer características de la cara y, una vez son normalizados, se emplea un clasificador SVM para su evaluación. La base de datos entrenada para este sistema fue *UADFV*, obteniendo un AUC del **89 %**.

DSP-FWA

Otro de los aspectos que se ha tenido en cuenta ha sido la resolución de los vídeos una vez manipulados. Li y Lyu [35] propusieron un sistema basado en CNN con el fin de detectar anomalías causadas por la manipulación en regiones faciales y sus alrededores. Este sistema fue entrenado por las bases de datos *UADFV* y *DeepFakeTIMIT*, logrando superar resultados anteriores. Posteriormente, estos mismos autores junto con otros [29] presentaron una versión mejorada de estos sistemas que designaron como *Dual Spatial Pyramid - Face Warp Artifacts* (DSP-FWA). En efecto, los resultados que se han conseguido han sido bastante buenos para las bases de datos entrenadas -alcanzando valores AUC cercanos al 100 %- tal y como se puede observar en la Tabla 2.2.

XceptionNet

XceptionNet [39] es una arquitectura basada en CNN, como resultado de una evolución del sistema Inception [40], anterior a ella.

- **Rössler *et al.* [4]**: en este estudio se desarrollaron cinco sistemas diferentes de detección enfocados a las técnicas de manipulación de *FaceForensics++* [4]. Estos cinco sistemas presentaban redes CNN basadas en diferentes características, no obstante, el sistema con mejor rendimiento fue la arquitectura *XceptionNet*. Este sistema fue re-entrenado con los

dos métodos de manipulación de *Identity Swap* de la base de datos de *FaceForensics++*: *DeepFakes* y *FaceSwap*, logrando resultados cercanos al 100 % en *accuracy* para los tres tipos de calidad que proporciona este estudio -*raw*, HQ y LQ-. Cuanto mayor es la calidad, mayor es la dificultad de poder manipular los vídeos, por lo tanto, menor esfuerzo para poder detectar los vídeos *fake*. Es por esta razón que los rendimientos son peores a medida que la calidad del vídeo baja, lo que supone un reto mayor para detectar vídeos en escenarios reales.

- **Dolhansky *et al.* [5]**: esta arquitectura también se utilizó para la detección en la base de datos *DFDC*, junto con una red neuronal profunda denominada *TamperNet*. En esta ocasión, *XceptionNet* fue entrenada con dos tipos de imágenes diferentes: la imagen entera de un *frame* o, por otro lado, recortando la imagen y mostrando únicamente la cara recortada. De este modo, se consiguieron mejores rendimientos con la arquitectura *XceptionNet* basada en imágenes donde solo se mostraba el rostro de la persona, con una precisión del **93 %** y un *recall* del **8,4 %**.

Capsule Forensic

Nguyen *et al.* presentaron en su trabajo un sistema basado en las recientes *Capsule Networks* [36]. Las redes CNN alcanzan grandes rendimientos pero requiere decenas de miles de imágenes para ello, además de que son incapaces de reconocer variaciones y deformaciones que puedan afectar a las mismas [41]. Por ello, se desarrollaron las denominadas *Capsule Networks*, que eran capaces de corregir estos errores. Sin embargo, esta técnica de detección -entrenada con bases de datos desconocidas- ofrece una mala capacidad de generalización frente al resto de bases de datos evaluadas, exceptuando *FaceForensics++* , que obtuvo un **96,6 %** en AUC.

DenseNet

Por último, Sabier *et al.* propusieron un sistema para detectar vídeos *fake* basados en el uso de información temporal de los mismos [37]. Para ello, utilizaron un tipo de red neuronal convolucional denominado *DenseNet*, donde todas las capas que lo forman están conectadas entre sí de manera que se retroalimentan en todo momento [42]. Además, este sistema incluye técnicas de alineamiento de la cara y una red neuronal recurrente (RNN) bidireccional. Las redes recurrentes almacenan información a modo de retroalimentación para las siguientes épocas. Esta arquitectura fue entrenada y evaluada con la base de datos *FaceForensics++*, superando los resultados de cualquier otro estudio anterior, con valores AUC en torno al **96 %**.

2.3. Conclusiones

En este capítulo se ha analizado el estado del arte de los diferentes tipos de manipulación y en concreto, las basadas en *Identity Swap*. Tras examinar con detalle este último tipo, se ha observado la evolución tanto visual como a nivel de detección de las dos generaciones de bases de datos, donde la 2ª de ellas ofrece un contenido *fake* que difícilmente es detectable al ojo humano.

Una vez se han estudiado los principales métodos de manipulación y detección de *Identity Swap*, se observa que las técnicas basadas en *Autoencoders* junto con las redes GAN ofrecen los mejores resultados visuales y más complicados de detectar, como es el caso de la base de datos *Celeb-DF* perteneciente a la 2ª generación. Por otro lado, los métodos de detección con mejor rendimiento han sido los que han proporcionado un estudio basado en redes neuronales profundas, como se observa en la Tabla 2.2.

Finalmente, se han considerado algunos aspectos a tener en cuenta respecto a las técnicas de detección estudiadas: *i)* análisis a nivel de imagen sin considerar las distintas regiones faciales para sistemas basados en *deep learning* y para cada una de las generaciones (a excepción del estudio realizado con el método *Visual Artifacts*), *ii)* las técnicas de manipulación examinadas en cada estudio suelen estar entrenadas con una base de datos únicamente (dos como mucho), y *iii)* no existe un protocolo uniforme entre los estudios -protocolos desconocidos y métricas diferentes-. Por estos motivos, el presente trabajo -al igual que en nuestro artículo publicado [9], proporciona un estudio que considera estos aspectos. En primer lugar, el estudio considera 2 métodos distintos: *i)* seleccionando la cara entera , y *ii)* seleccionando las regiones faciales que conforman la cara como los ojos o la boca, entre otros. En segundo lugar, los sistemas propuestos han sido entrenados y evaluados con cada base de datos y se ha utilizado un **protocolo uniforme** para todos ellos, con el fin de obtener unos resultados equiparables entre sí. Además, se ha medido la capacidad de generalización que tienen los modelos frente a bases de datos no entrenadas.

3

Sistemas Propuestos

En este capítulo se muestran los sistemas utilizados para la detección de vídeos *fake* a nivel de imagen para el tipo de manipulación *Identity Swap*, considerando 4 bases de datos -*UADFV*, *FaceForensics++*, *Celeb-DF* y *DFDC*- pertenecientes a las 2 generaciones ya comentadas. En la siguiente sección se presenta finalmente el método propuesto para los sistemas de detección *fake*.

3.1. Método Propuesto

La arquitectura desarrollada para la detección de vídeos *fake* consta de dos módulos independientes, tal y como se muestra en la Figura 3.1.

Por un lado, todos los vídeos -tanto reales como *fake*- son enviados a un primer módulo de **Segmentación Facial por Regiones**. Este detector facial se encarga de identificar la cara en todos los *frames* de cada vídeo, y una vez son identificadas, se separan por regiones faciales: cara entera, ojos, nariz, boca y resto de regiones. Por lo tanto, se han considerado dos métodos: *i)* seleccionando la cara entera como entrada para el sistema de detección, y *ii)* seleccionando regiones específicas -como los ojos o la boca, entre otros- como entrada para el sistema de detección. Por otro lado, una vez se ha segmentado cada región, son enviados a los **Sistemas de Detección *Fake*** -cada uno a su modelo correspondiente-, donde son entrenados o evaluados para la posterior detección. Como se ha comentado en los capítulos previos, el análisis se realiza a nivel de imagen por cada región y base de datos evaluada.

3.2. Segmentación Facial por Regiones

En primer lugar, todos los vídeos son procesados *frame* a *frame* y enviados a este sistema de segmentación facial. Mediante la herramienta ***OpenFace*** [10], se extraen 68 *landmarks* o puntos de referencia faciales como se observa en la Figura 3.2. Cabe destacar que esta herramienta es robusta frente a variaciones en la pose, distancia a la cámara y condiciones de luz, por lo que ofrece fiabilidad a la hora de extraer estos puntos. El siguiente paso a realizar corresponde con el proceso principal de este sistema, ya que supone la **segmentación facial por regiones**. Como se ha comentado previamente, se ha considerado el uso de dos métodos diferentes a la hora de detectar vídeos.

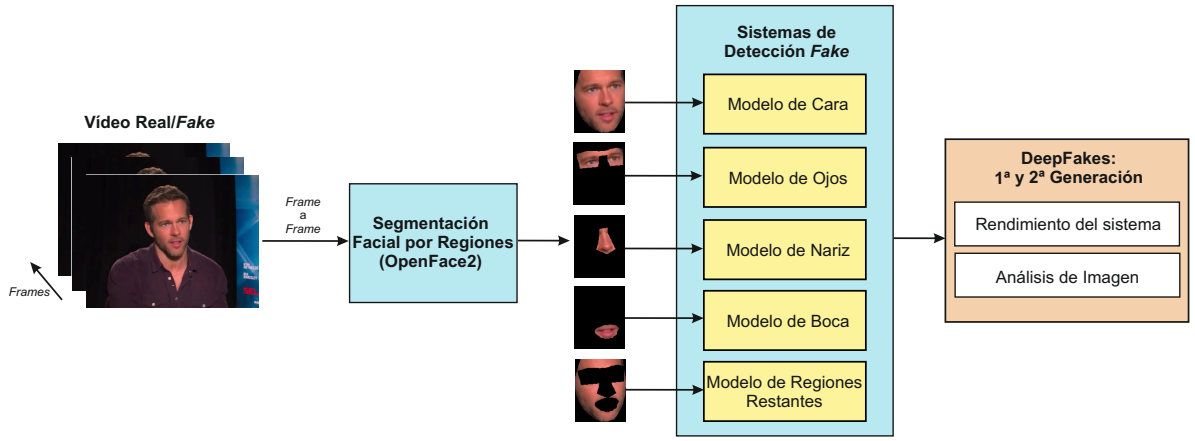


Figura 3.1: Arquitectura del sistema desarrollado en este trabajo.

Modelo de Cara Completa

El primer método está formado por un único modelo correspondiente a **toda** la **cara**. Para ello, se han tomado todos los puntos que conforman la mandíbula y el contorno exterior, así como los puntos que forman las cejas: del 1 al 17 y del 18 al 27, respectivamente. Estos *landmarks* se han unido formando un polígono irregular, donde finalmente se ha elevado la zona de las cejas para que la frente formara parte de este polígono, tal y como se observa en la Figura 3.2 (polígono formado por las líneas amarillas exteriores).

Modelos por Regiones Faciales

El segundo método está formado por cuatro aproximaciones diferentes correspondientes a las siguientes **regiones faciales**:

- **Ojos**: usando los puntos de las cejas -del 18 al 27- para la parte superior y los puntos 1, 2, 16 y 17 para la parte inferior, se forma un polígono con forma de antifaz, excluyendo las zonas que pertenecen a la nariz.
- **Nariz**: se han tenido en cuenta los puntos 22 y 23 para delimitar el inicio de la nariz con las cejas, al igual que los puntos 40 y 43 para definir el ancho. Por otro lado, se han usado los puntos del 28 al 36 correspondientes a la nariz y se ha ajustado para obtener un polígono que coincida con la zona nasal.
- **Boca**: se han tomado los *landmarks* 49, 51-53, 55 y 57-59 referentes a la boca para formar un polígono con forma de elipse. Finalmente, se ha aumentado el tamaño de este polígono en un 15 %.
- **Resto**: este modelo se extrae tras eliminar los modelos de los ojos, de la nariz y de la boca, obteniendo un polígono del resto de las regiones faciales -frente, mejillas y toda la zona de la mandíbula-.

Es importante remarcar que para la obtención de cada uno de los modelos faciales, en primer lugar se ha creado el polígono de cada región para posteriormente eliminar las zonas no pertenecientes a ella. Este último proceso se ha conseguido pintando de negro todas las zonas excluidas (véase Figura 3.1).

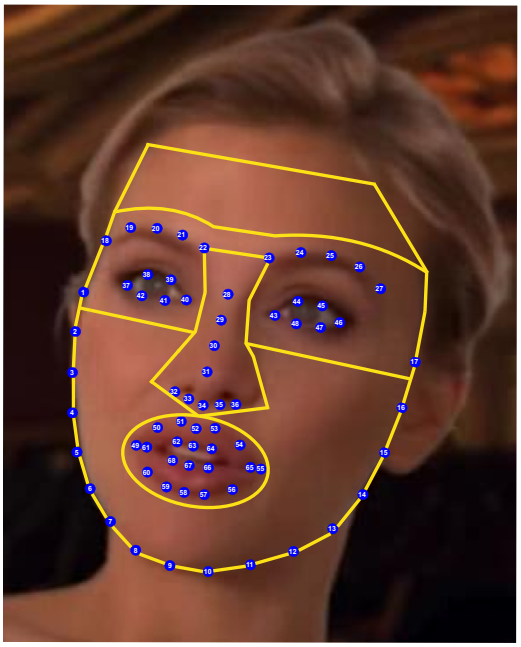


Figura 3.2: Ejemplo de las diferentes regiones faciales (ojos, boca, nariz y resto) usando los 68 *landmarks* extraídos a partir de la herramienta *OpenFace*.

3.3. Sistemas de Detección *Fake*

Para este estudio, se han considerado tres **Sistemas de Detección *Fake*** en el estado del arte: dos de ellos están basados en CNN y el restante basado en *Capsule Networks* -aunque también haga uso en parte de CNN-. Esta selección se debe a los rendimientos tan altos que han obtenido estas redes neuronales en los estudios previos, como se comenta en el Capítulo 2.2.3. De este modo, cada modelo sigue un proceso de entrenamiento y evaluación, donde finalmente las imágenes son clasificadas como *real* o *fake*.

Para los tres sistemas se han considerado las siguientes características:

- El número de épocas de entrenamiento para cada base de datos ha sido de 5 para *FaceForensics++*, *Celeb-DF* y *DFDC* y 20 épocas para *UADFV*. Esto se debe principalmente a que las primeras bases de datos contienen una cantidad de imágenes de entrenamiento tan alta que no necesitan muchas épocas para obtener su mejor rendimiento.
- El modelo escogido para la evaluación final corresponde con el que obtiene mejor rendimiento basado en el *accuracy* en validación.
- Para la función de coste se ha utilizado el optimizador *Adam* (*learning rate* = 0'002, $\beta_1 = 0'9, \beta_2 = 0'999$) basado en entropía cruzada binaria (*binary cross-entropy*).
- *Batch size*: 32.
- Dimensiones de las imágenes (*input* de las redes CNN): 200x200x3.
- Uso de la técnica *data augmentation* para una mejora del rendimiento y evitar *overfitting*.
- Utilización de una tarjeta gráfica *NVIDIA GeForce RTX 2080 Ti GPU* para todos los experimentos.

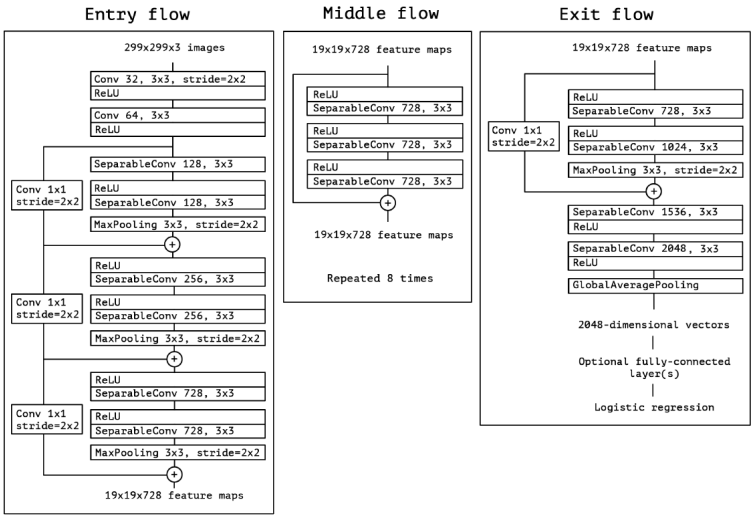


Figura 3.3: Arquitectura de la red *XceptionNet*. Fuente: [39].

3.3.1. *XceptionNet*

Las CNN son un tipo de red neuronal especializadas en tratamiento de imágenes. En este caso, ***XceptionNet*** es una red neuronal inspirada en *Inception* [40], un tipo de red anterior a esta. La arquitectura está formada por 36 capas convolucionales repartidas en 14 módulos diferentes, como se puede observar en la Figura 3.3. Además, esta arquitectura obtiene mejores rendimientos en comparación con otras redes neuronales como *Inception*, *VGG* o *ResNet* [39].

Para el desarrollo de este primer sistema se ha utilizado la librería de *Keras* usando *Tensorflow* como *backend* y parte del código implementado procede del libro *Deep Learning with Python*[1] [11]. Por otra parte, se ha seguido un procedimiento muy similar al desarrollado en [4]:

1. Se ha utilizado inicialmente una red pre-entrenada con la base de datos *ImageNet* [43], que cuenta con 1000 clases diferentes de imágenes -desde todo tipo de animales y seres vivos hasta objetos cotidianos y diferentes deportes, entre otros muchos-.
2. Esta última capa *fully-connected* (que cuenta con 1000 clases diferentes) se ha cambiado por una nueva con dos clases -*real* y *fake*-.
3. Todos los pesos de las capas se han congelado exceptuando esta última capa *fully-connected* añadida -se encarga de clasificar las imágenes en *real* o *fake*-, la cual se ha entrenado unas pocas épocas.
4. Finalmente, se han descongelado todas las capas de la red y se ha entrenado unas épocas más.

3.3.2. *Capsule Forensics*

Este método denominado ***Capsule Forensics*** utiliza *Capsule Networks*, la cual ofrece una arquitectura más novedosa frente a las redes neuronales convolucionales [36]. Esta red está compuesta por cápsulas formadas a su vez por redes de neuronas cuya salida representa una propiedad diferente de una misma característica. De este modo, estas propiedades permiten a la cápsula aprender partes de una imagen junto con las características que cada una de estas partes aporta. Además, la entrada de una cápsula es la salida de una CNN [41].

[1] https://github.com/fchollet/deep-learning-with-python-notebooks

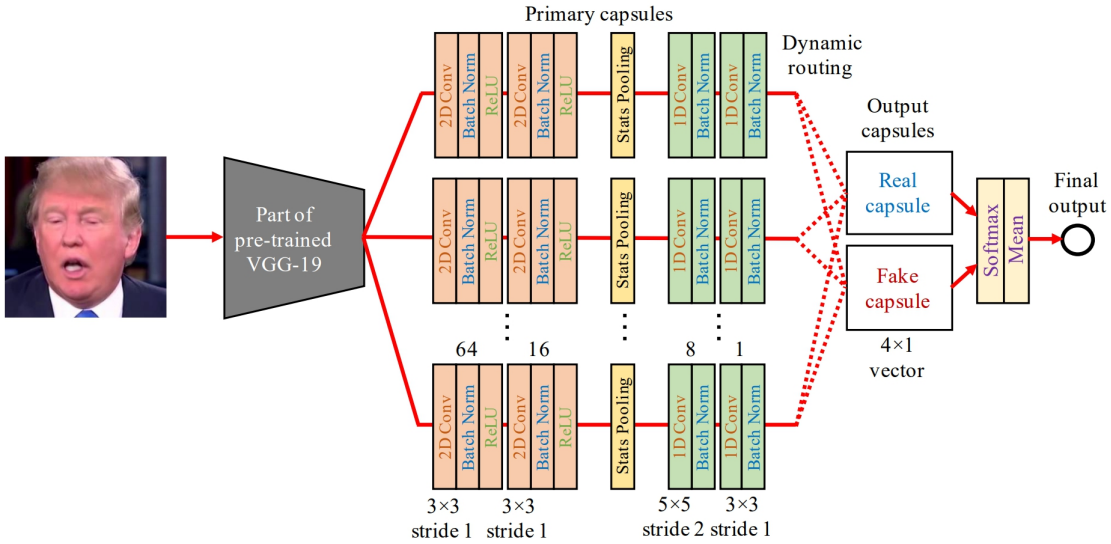


Figura 3.4: Arquitectura de *Capsule Forensics*. Fuente: [36].

Como se observa en la Figura 3.4, la arquitectura *Capsule Forensics* consta de:

- Una red neuronal convolucional *VGG* con 19 capas, previamente entrenada con *ImageNet* a la que se envían las imágenes, como extractor de características.
- Una *Capsule Network* que incluyen 10 cápsulas primarias (*Primary Capsules*) y dos cápsulas de salida (*Output Capsules) - real* y *fake*-. La salida de la red convolucional se envía a las 10 cápsulas primarias.
- Un algoritmo denominado *dynamic routing*, encargado de calcular la concordancia entre las características extraídas por las cápsulas primarias, para posteriormente enviar los resultados a la cápsula de salida más adecuada (*real* o *fake*). En último lugar, se calcula la clase final.

La implementación de este sistema se ha llevado a cabo con la librería *PyTorch* y el código proporcionado por los autores disponible en la plataforma *GitHub*[2].

3.3.3. DSP-FWA

Este último método denominado ***Dual Spatial Pyramid for Exposing Face Warp Artifacts*** es una versión mejorada basada en FWA [35] que incluye un módulo de *Spatial Pyramid Pooling (SPP)*, cuyo objetivo reside en poder tratar imágenes -en este caso caras- con diferentes resoluciones.

La arquitectura del *DSP-FWA* está constituida por 3 módulos conectados entre sí, como se muestra en la Figura 3.5:

- El primer módulo está formado por una CNN denominada ***ResNet***. Esta red, al igual que *XceptionNet* está pre-entrenada con *ImageNet* y está compuesta por 50 capas de *Residual Networks* [44], del mismo modo que propusieron los autores de este sistema [35].

[2]https://github.com/nii-yamagishilab/Capsule-Forensics-v2

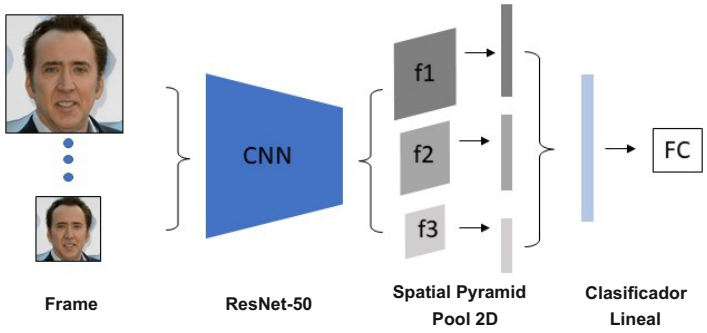


Figura 3.5: Arquitectura del sistema *DSP-FWA*[3].

- Seguidamente, el segundo módulo está formado por una red neuronal denominada ***Spatial Pyramid Pool 2D*** (SPPNet), la cual puede tratar imágenes con cualquier resolución [45].
- Finalmente, el tercer módulo corresponde con el clasificador de este sistema. En este caso, se utiliza un **Clasificador Lineal** (*Linear Classifier*) con dos clases para diferenciar entre una imagen *real* o una *fake*.

Al igual que el anterior sistema, la implementación se ha llevado a cabo con la librería *PyTorch* y el código proporcionado por los autores disponible en la plataforma *GitHub*[3].

[3]https://github.com/danmohaha/DSP-FWA

4

Desarrollo Experimental

En este capítulo se muestra el protocolo experimental que se ha llevado a cabo para las bases de datos estudiadas. Tras evaluar los modelos, se han analizado los rendimientos obtenidos y comparado los sistemas por regiones faciales. Seguidamente, se ha mostrado mediante mapas de calor las zonas de la cara donde las redes de los sistemas se fijan a la hora de tomar una decisión. Finalmente, se ha medido la capacidad de generalización que tiene cada sistema frente a bases de datos no utilizadas durante el entrenamiento.

4.1. Protocolo Experimental

Las bases de datos se han analizado independientemente, obteniendo 5 *datasets* por cada base de datos (uno por cada región facial a estudiar). A su vez, cada conjunto de datos se ha divido en un conjunto de desarrollo (alrededor del **80 %** de las identidades) y un conjunto de evaluación final (alrededor del **20 %** de las identidades restantes) no utilizado durante el proceso de entrenamiento de los sistemas. Asimismo, el conjunto de datos de desarrollo se ha divido en entrenamiento (alrededor del 80 % de las identidades del conjunto de desarrollo) y en validación (alrededor del 20 %). Además, el análisis se ha realizado a nivel de imagen y se han extraído una media de 300 *frames* por cada vídeo, seleccionando 270 *frames* para la fase de entrenamiento y 100 para las fases de validación y evaluación, tal y como se considera en [4]. Particularmente, para la base de datos *DFDC* se han seleccionado 50 *frames* por vídeo debido a la dificultad de detectar los 68 *landmarks* de manera correcta en estos escenarios no controlados.

A diferencia de algunos estudios previos, estos *datasets* están formados por el mismo número de vídeos reales como de vídeos *fake*. Este aspecto es importante a destacar ya que ofrece una evaluación más equilibrada e imparcial por un lado, y una capacidad de generalización del sistema de detección *fake* frente a identidades no vistas, por otro lado.

4.1.1. *UADFV*

La base de datos *UADFV*, como se ha comentado en el Capítulo 2.2.1, consta de **49** vídeos reales y **49** vídeos *fake* y **se han tomado las mismas decisiones para ambas clases a la hora de dividir los conjuntos de datos**:

	Nº Identidades	Vídeos	
		Reales	Fakes
Grupo 1	29	295	3962
Grupo 2	4	40	86
Grupo 3	5	54	107
Grupo 4	10	91	723
Grupo 5	10	100	728
Grupo 6	1	10	33
Total	**59**	**590**	**5639**

Tabla 4.1: Estructura de los diferentes grupos entre identidades.

- **Desarrollo**: 38 vídeos (77,55 %).
  - **Entrenamiento:** 30 vídeos (78,95 %).
  - **Validación:** 8 vídeos (21,05 %).
- **Evaluación:** 11 vídeos que corresponden con la identidad de Donald Trump (22,45 %).

4.1.2. *FaceForensics++*

Para *FaceForensics++*, la técnica de manipulación para el análisis de este trabajo ha sido *FaceSwap* y **se ha seleccionado el mismo protocolo considerado en [4] para ambas clases**. Esta base de datos cuenta con **1000** vídeos reales y **1000** vídeos *fake*:

- **Desarrollo:** 860 vídeos (86 %).
  - **Entrenamiento:** 720 vídeos (83,72 %).
  - **Validación:** 140 vídeos (16,28 %).
- **Evaluación:** 140 vídeos (14 %).

4.1.3. *Celeb-DF*

Para *Celeb-DF*, se ha realizado un estudio previo para obtener los diferentes *datasets* en los que se ha dividido. Es importante recordar que hay un total de **890** vídeos reales procedentes de YouTube (**590** corresponden con las **59** identidades que conforman los vídeos *fake* y 300 restantes) y **5639** vídeos *fake*. Estos últimos vídeos se han generado intercambiando las caras de las identidades entre sí, organizándolas en grupos según si ha habido intercambio o no. De esta manera, se han creado 6 grupos diferentes atendiendo a las características descritas en la Tabla 4.1.

Una vez se han agrupado las diferentes identidades, se ha definido un protocolo lo más equilibrado posible. Por otra parte, la restricción del primer grupo (reúne más de la mitad de los vídeos totales) al no poder dividirse en los *datasets* de desarrollo y evaluación, ha propiciado que este grupo se asigne al conjunto de desarrollo. Además, se han añadido a este conjunto de datos de desarrollo los grupos 4 y 6. Por último, los grupos correspondientes al *dataset* de evaluación han sido el 2, 3 y 5.

De igual modo, se han añadido al conjunto de desarrollo de vídeos reales, los 300 que pertenecían a *Celeb-DF* (procedentes de YouTube sin identidades conocidas) y los 1000 vídeos pertenecientes a la base de datos *FaceForensics++* (también de YouTube).

Por lo tanto, el número de vídeos reales asciende a **1890**:

- **Desarrollo:** 1696 vídeos (89,74 %).
  - **Entrenamiento:** 1357 vídeos (80 %).
  - **Validación:** 339 vídeos (20 %).
- **Evaluación:** 194 vídeos (10,26 %).

Por otro lado, de los 5639 vídeos *fake* pertenecientes a *Celeb-DF*, se han eliminado 8 de ellos debido a que se intercambiaban identidades entre los dos conjuntos de datos, obteniendo un total de **5631** vídeos:

- **Desarrollo:** 4710 vídeos (83,64 %).
  - **Entrenamiento:** 3768 vídeos (80 %).
  - **Validación:** 942 vídeos (20 %).
- **Evaluación:** 921 vídeos (16,36 %).

Finalmente, **como el número de vídeos *fake* en relación con los vídeos reales** sigue siendo muy superior (5631 frente a 1890), **la red es entrenada con el mismo número de imágenes de ambas clases en cada época hasta que se hayan examinado todas ellas**.

4.1.4. *DFDC*

Para la base de datos *DFDC*, se ha utilizado el mismo protocolo que los autores proponen [5]. Por un lado, se han contabilizado **1131** vídeos reales:

- **Desarrollo:** 855 vídeos(75,6 %).
  - **Entrenamiento:** 684 vídeos (80 %).
  - **Validación:** 171 vídeos (20 %).
- **Evaluación:** 276 vídeos (24,4 %).

Por otro lado, se han generado **4119** vídeos *fake*:

- **Desarrollo:** 3615 vídeos(87,76 %).
  - **Entrenamiento:** 2892 vídeos (80 %).
  - **Validación:** 723 vídeos (20 %).
- **Evaluación:** 504 vídeos (12,24 %).

Del mismo modo que en *Celeb-DF*, **el número de vídeos *fake* es muy superior al de vídeos reales**, por lo que **la red es entrenada con el mismo número de imágenes de ambas clases en cada época hasta que se hayan examinado todas ellas**.

4.2. Resultados Experimentales

En esta sección se analizan los resultados que se han obtenido en los diferentes experimentos realizados, así como la capacidad de generalización que tiene cada modelo frente a bases de datos diferentes a la entrenada.

Tabla 4.2: Rendimientos obtenidos en evaluación para las bases de datos correspondientes a la **1ª generación**. La primera tabla corresponde con rendimientos en términos de AUC, mientras que la segunda en términos de EER. Los mejores resultados conseguidos para cada modelo se señalan en **negrita** y en azul y naranja las regiones faciales que proporcionan el mejor y el peor resultado, respectivamente.

	UADFV			***FaceForensics++***		
AUC (%)	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA
Cara	**100**	99,90	**99,74**	**99,40**	**99,52**	**99,48**
Ojos	99,70	**100**	98,76	92,70	95,32	92,07
Nariz	94,70	99,30	96,92	86,30	90,09	85,21
Boca	95,40	99,56	90,88	93,90	96,18	94,99
Resto	97,30	94,83	91,99	85,50	86,61	86,81
EER (%)						
Cara	**1,00**	2,00	**1,00**	**3,31**	**2,75**	**3,35**
Ojos	2,20	**0,28**	6,07	14,23	10,29	15,02
Nariz	13,50	3,92	11,50	21,97	17,51	22,49
Boca	12,50	3,20	11,84	13,77	9,66	13,25
Resto	7,90	12,30	16,20	22,37	21,58	21,24

4.2.1. Análisis del Rendimiento por Regiones Faciales

Una vez se ha terminado la fase de desarrollo, se seleccionan los modelos con el mejor rendimiento basado en validación para cada región facial. Posteriormente, se evalúa cada modelo con el conjunto de datos de evaluación final no utilizado durante el entrenamiento, para las dos aproximaciones descritas en la Sección 3.2: *i)* seleccionando la cara completa (Cara), y *ii)* seleccionando regiones faciales específicas. Para este estudio, las métricas escogidas han sido *Area Under the Curve*(**AUC**) y *Equal Error Rate*(**ERR**), las cuales se han comentado previamente en el Capítulo 2.2.3.

1ª Generación

En la Tabla 4.2 se muestran los rendimientos obtenidos en evaluación para las bases de datos de la 1ª generación. Para cada base de datos y sistema de detección, los mejores resultados conseguidos se indican **negrita** y en azul y naranja las regiones faciales que proporcionan el mejor y el peor resultado, respectivamente.

En primer lugar, el análisis se centra en los rendimientos obtenidos para el modelo de la cara entera (Cara), con resultados AUC muy cercanos al **100 %** para ambas bases de datos. Exceptuando el sistema *Capsule Forensics* para *UADFV*, el resto de sistemas alcanzan el mejor rendimiento en comparación con los modelos restantes, además de que los valores EER no pasan del **4 %**.

En segundo lugar, los modelos de las distintas regiones faciales alcanzan diferentes rendimientos según se hayan generado los vídeos *fake* para cada base de datos evaluada. Por un lado, los mejores resultados que se alcanzan en *UADFV* son para el modelo de los ojos, con un valor del **100 %** de AUC y un **0,28 %** de EER para el sistema *Capsule Forensics*. Esto se debe principalmente a que en la fase de entrenamiento del modelo solo hay una identidad en los vídeos *fake*, por lo que los sistemas al identificar los ojos de esa persona, detectan con facilidad este contenido falso. No obstante, los sistemas alcanzan resultados muy dispares para el resto de regiones faciales debido a las características de cada una: mientras que el sistema basado en

XceptionNet obtiene un **97,30 %** de AUC y un **7,90 %** de EER para el modelo del resto de la cara, el sistema *Capsule Forensics* logra rendimientos muy buenos para los modelos de la nariz y de la boca, con valores AUC cercanos al **100 %** y valores por debajo del **4 %** de EER.

Por otro lado, para *FaceForensics++* los mejores rendimientos se obtienen para el modelo de la boca, con valores entre el **93 %** y el **96 %** para AUC y entre el **9 %** y el **13 %** para EER. La falta de calidad en los dientes -se muestran borrosos frente a los vídeos originales- y las inconsistencias en los labios han hecho que los sistemas detecten con mayor facilidad este modelo. De manera similar ocurre con el modelo de los ojos, logrando alcanzar un **95,32 %** de AUC y un **10,29 %** de EER para el sistema *Capsule Forensics*. Por otro lado, los peores resultados corresponden con el modelo del resto de la cara para los sistemas *XceptionNet* y *Capsule Forensics* y el modelo de la nariz para DSP-FWA. En este caso, los rendimientos se encuentran por debajo del **87 %** de AUC y por encima del **20 %** de EER.

Finalmente, el análisis de los rendimientos muestra una evolución entre las bases de datos, donde se observa una mejora en la última de ellas, *FaceForensics++*. Como se detalla en la Tabla 4.2, los sistemas detectan con más facilidad los vídeos *fake* procedentes de la base de datos *UADFV*: por un lado, esto se debe principalmente a que los vídeos *fake* se han generado con una única identidad impostora (Nicolas Cage); por otro lado, las mejoras en el contraste de color o la máscara han propiciado unos peores resultados en *FaceForensics++*. Además, en esta primera generación cabe destacar que el sistema *Capsule Forensics* ha obtenido los mejores resultados frente a los dos restantes en la mayoría de los modelos para ambas bases de datos, llegando a alcanzar rendimientos superiores al **95 %** de AUC en la mayoría de las regiones faciales evaluadas.

2ª Generación

Para la 2ª generación se observa una mejora en la calidad y realismo de los vídeos *fake* y consecuentemente, unos rendimientos en detección de *fake* inferiores a los obtenidos en la 1º generación, tal y como se muestra en la Tabla 4.3.

De forma similar a la 1ª generación, los sistemas ofrecen mejores rendimientos en los modelos correspondientes a la cara entera para ambas bases de datos. Sin embargo, estos resultados son muy dispares: para la bases de datos *Celeb-DF*, se han logrado alcanzar valores por encima del **96 %** de AUC y en torno al **6 %** de EER. Por otra parte, los rendimientos logrados para *DFDC* son inferiores, con un **91 %** de AUC y un **17,55 %** de EER para el sistema *XceptionNet* y valores en torno al **87 %** de AUC y al **22 %** de EER para los sistemas *Capsule Forensics* y *DSP-FWA*. Por lo tanto, esta última base de datos representa un mayor desafío a la hora de identificar contenido falso, siendo una de las posibles razones la menor cantidad de datos de entrenamiento para *DFDC*.

Para los modelos correspondientes a las distintas regiones faciales, el modelo de los ojos logra los mejores resultados. Para *Celeb-DF*, el sistema *XceptionNet* alcanza el mejor rendimiento frente a los dos sistemas restantes, con un **95,20 %** y un **11,11 %** de AUC y EER, respectivamente. Ese mismo sistema consigue obtener también el mejor rendimiento para *DFDC*, con un **83,90 %** de AUC y un **23,82 %** de EER.

Por el contrario, los modelos con peores resultados corresponden a la nariz y al resto de la cara. Por un lado, los sistemas *XceptionNet* y *DSP-FWA*, con un **78 %** de AUC y en torno al **29 %** de EER, ofrecen el peor rendimiento para el modelo de la nariz en *Celeb-DF*, mientras que el sistema *Capsule Forensics* obtiene un resultado algo superior para el resto de la cara, con un **78,20 %** de AUC y un **26,93 %** de EER. Por otro lado, para la base de datos *DFDC* los 3 sistemas ofrecen peores rendimientos en el modelo del resto de la cara, con resultados de AUC entre **65 %** y **76 %** y de EER entre el **29 %** y el **38 %**. De este modo, se ha mejorado

Tabla 4.3: Rendimientos obtenidos en evaluación para las bases de datos correspondientes a la **2ª generación**. La primera tabla corresponde con rendimientos en términos de AUC, mientras que la segunda en términos de EER. Los mejores resultados conseguidos para cada modelo se señalan en **negrita** y en azul y naranja las regiones faciales que proporcionan el mejor y el peor resultado, respectivamente.

	Celeb-DF			***DFDC***		
AUC (%)	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA
Cara	**98,30**	**96,80**	**98,60**	**91,00**	**87,45**	**87,94**
Ojos	95,20	91,10	93,80	83,90	83,12	79,68
Nariz	78,00	80,20	78,00	81,50	81,50	74,91
Boca	84,80	82,10	83,80	79,50	78,14	77,96
Resto	84,40	78,20	84,00	76,50	72,42	65,04
EER (%)						
Cara	**6,41**	**7,98**	**5,98**	**17,55**	**21,39**	**22,37**
Ojos	11,11	15,26	12,75	23,82	25,06	26,60
Nariz	29,05	26,82	28,95	26,80	26,53	32,33
Boca	23,49	24,53	23,78	27,59	27,92	28,97
Resto	23,48	26,93	23,74	29,94	32,56	38,21

notablemente la adaptación de las máscaras *fake* a la cara original, eliminando cualquier tipo de visibilidad de esta cara.

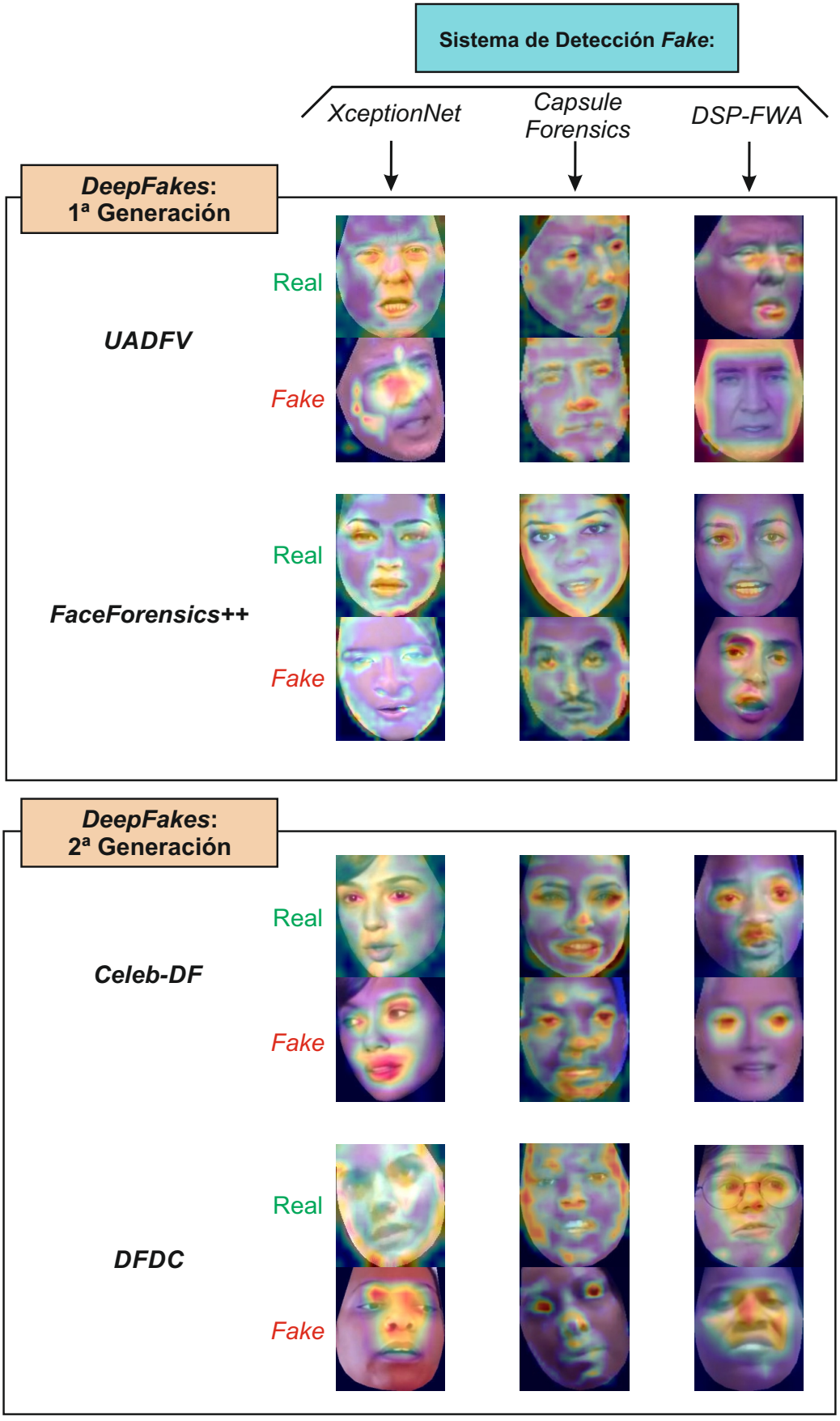
Por lo tanto, la base de datos ***DFDC*** obtiene los peores rendimientos de la 2ª generación y en consecuencia, es la que contiene los vídeos *fake* más difíciles a detectar por los sistemas evaluados de las dos generaciones **a nivel de imagen**. Asimismo, el sistema que ofrece mejores resultados para ambas bases de datos es *XceptionNet* en lugar de *Capsule Forensics*, como ocurría para la 1ª generación. En conclusión, las técnicas de manipulación han mejorado notablemente desde la 1ª generación hasta la 2ª, impidiendo tanto al ojo humano como a los sistemas de detección *fake* identificar este contenido. Cabe destacar que para ambas generaciones, los sistemas obtienen los mejores rendimientos en el modelo correspondiente con la **cara entera** en la mayoría de los casos. Además, estos sistemas logran peores rendimientos ante las bases de datos de la última generación y como se ha comprobado, no hay un sistema de detección dominante: en función de la generación y modelo facial, unos sistemas obtienen mejores resultados que otros. No obstante, pese a que haya habido una evolución notable entre una generación y otra (véase Sección 2.2.1), aún hay aspectos a mejorar aunque para el ojo humano esta tarea sea imperceptible. Como se observa en la Tabla 4.3, estos aspectos tanto en la base de datos *Celeb-DF* como *DFDC* corresponden con la zona de los ojos, debido a que los modelos de esta región facial obtienen los mejores rendimientos en la identificación de vídeos *fake*.

Mapas de Calor *Grad-CAM*

Como se ha comentado anteriormente, cada sistema de detección *fake* está formado por una arquitectura independiente al resto, lo que ha proporcionado unos rendimientos diferentes para un mismo modelo facial. Por ello, se han utilizado técnicas de visualización a través de mapas de calor con el objetivo de visibilizar aquellas zonas donde el sistema toma la decisión final -*real* o *fake*-.

En este caso, se ha empleado la técnica *Gradient-weighted Class Activation Mapping* (*Grad-CAM*), la cual utiliza la información del gradiente de la última capa de una CNN para discriminar las zonas donde la red se fija a la hora de tomar una decisión [46]. Por lo tanto, mediante este

Figura 4.1: Ejemplos de aciertos del sistema para imágenes reales y *fake* tratadas con mapas de calor *Grad-CAM* mediante el modelo facial del rostro entero (Cara) para los tres sistemas implementados. Las imágenes se han extraído de las bases de datos analizadas en este estudio.



mapa de calor se observa las regiones faciales que se han tenido en cuenta, lo que permite compararlo con los rendimientos alcanzados en los modelos faciales.

Tal y como se muestra en la Figura 4.1, las zonas que están marcadas con tonos más intensos (rojo y amarillo) corresponden con aquellas partes en donde la red se ha fijado más, mientras que las zonas con tonos más apagados (azul, verde o el color original), coincide con las que ha tenido menos en cuenta. Como se puede observar, se han tomado ejemplos de aciertos de las 4 bases de datos analizadas y para cada uno de los sistemas de detección *fake* [1,2] y se ha escogido el modelo de la **región entera** (cara completa).

Para la mayoría de las bases de datos y sistemas analizados, las CNN se fijan en la región central de la cara, resaltando notablemente la zona de los **ojos**. Al comparar estos resultados junto con los rendimientos obtenidos, se observa que el mejor modelo por regiones faciales ha sido el de los ojos, lo que indica que esta región es imprescindible para los sistemas: aunque se haya seleccionado el modelo de la cara completa para este estudio, las diferentes redes han tenido presente esta zona para la decisión final, lo que confirma, por un lado, la habilidad del modelo de los ojos para detectar imágenes *fake* y por otro lado, la debilidad de las técnicas de manipulación en esta región al tratar con estas imágenes, como se ha comentado anteriormente.

Otras regiones también destacadas por estos mapas de calor han sido la nariz y la boca: la borrosidad en los dientes o la baja calidad de los labios y la nariz han ayudado a que los sistemas tengan en cuenta estas zonas a la hora de identificar una imagen como *fake*. En bases de datos como *Celeb-DF*, estos aspectos han sido significativos para los tres sistemas, otros en cambio, como es el caso de *FaceForensics++*, en el que los sistemas han tenido diferentes interpretaciones: de forma similar a *Celeb-DF*, el sistema DSP-FWA se ha fijado en la zona central de la cara, mientras que para el resto de sistemas *XceptionNet* y *Capsule Forensics*, se ha tenido en cuenta estas regiones junto con los límites de la máscara facial. De igual manera, el sistema DSP-FWA para la base de datos *UADFV* ha considerado únicamente las zonas exteriores de la máscara -es decir, la cara original- para las imágenes *fake* y la zona de los ojos y la boca para las imágenes reales.

Como se ha comprobado, aunque las técnicas de manipulación digital hayan evolucionado notablemente, hoy en día algunas regiones faciales ayudan en la labor de la detección, como es el caso de los ojos.

4.2.2. Capacidad de Generalización

En esta sección se analiza la capacidad de generalización que tienen los sistemas frente a bases de datos no utilizadas durante el entrenamiento. En este caso, se ha seleccionado el modelo correspondiente a la cara completa, debido a que ha proporcionado los mejores rendimientos.

En la Tabla 4.4 se muestran los resultados obtenidos en términos de AUC(%). Cada modelo entrenado, por lo tanto, se evalúa con el resto de bases de datos. Los mejores resultados logrados por cada modelo se señalan en **negrita** y en azul y naranja, los que proporcionan el mejor y el peor rendimiento con el resto de bases de datos, respectivamente.

Como se detalla en la tabla, para los tres sistemas se cumple que los modelos entrenados con su propia base de datos obtienen los mejores rendimientos en la evaluación, tal y como se ha estudiado previamente. Por otra parte, el análisis ofrece patrones característicos entre sistemas (sin tener en cuenta la propia base de datos a la que pertenece):

- Para los tres sistemas se obtienen:

[1] https://github.com/fchollet/deep-learning-with-python-notebooks/
[2] https://github.com/jacobgil/pytorch-grad-cam

Tabla 4.4: Tabla de capacidad de generalización de los modelos correspondientes a la cara entera, evaluados con las bases de datos estudiadas en este trabajo. Los mejores resultados logrados por cada modelo se señalan en **negrita** y en azul y naranja, los que proporcionan el mejor y el peor rendimiento con el resto de bases de datos, respectivamente. Todos los valores se muestran en términos de AUC(%).

	XceptionNet	**Evaluación**			
		UADFV	*FaceForensics++*	*Celeb-DF*	*DFDC*
Entrenamiento	*UADFV*	**100**	52,00	66,58	68,41
	FaceForensics++	79,52	**99,40**	53,62	45,05
	Celeb-DF	94,75	50,00	**98,30**	72,02
	DFDC	78,21	41,02	81,04	**91,17**

	Capsule Forensics	**Evaluación**			
		UADFV	*FaceForensics++*	***Celeb-DF***	*DFDC*
Entrenamiento	*UADFV*	**99,90**	61,82	61,55	63,27
	FaceForensics++	60,25	**99,52**	54,76	36,60
	Celeb-DF	96,73	59,92	**96,80**	70,89
	DFDC	86,32	40,49	77,56	**87,45**

	DSP-FWA	**Evaluación**			
		UADFV	*FaceForensics++*	*Celeb-DF*	*DFDC*
Entrenamiento	*UADFV*	**99,74**	54,58	59,08	67,12
	FaceForensics++	89,06	**99,48**	58,87	55,84
	Celeb-DF	98,21	46,28	**98,59**	77,65
	DFDC	85,81	45,84	78,50	**87,94**

  - Los rendimientos más altos para los modelos de *FaceForensics++* y *Celeb-DF* evaluando la base de datos *UADFV*, con valores superiores al **94 %** de AUC para el modelo de la 2ª generación.
  - Los mejores rendimientos para el modelo de *UADFV* evaluando la base de datos *DFDC*, aunque por debajo del **70 %** de AUC.
  - Los peores rendimientos para los modelos de *Celeb-DF* y *DFDC* evaluando la base de datos *FaceForensics++*, con valores de AUC entre el **40 %** y el **60 %**.
  - Los rendimientos más bajos para el modelo de *FaceForensics++* evaluando la base de datos *DFDC*, con valores de AUC que varían entre el **36 %** y el **56 %**.

- Para los sistemas *Capsule Forensics* y DSP-FWA se alcanzan los mejores rendimientos en los modelos de *DFDC* al evaluar la base de datos de *UADFV*, con valores en torno al **86 %** de AUC.

- Para los sistemas *XceptionNet* y DSP-FWA, se obtiene los rendimientos más bajos en los modelos de *UADFV* al evaluar la base de datos *FaceForensics++*, por debajo del **55 %** de AUC.

En general, los modelos detectan con mayor facilidad los vídeos *fake* procedentes de la base de datos ***UADFV*** debido a que, por un lado, esta base de datos fue la primera en aparecer y contiene numerosos defectos que ayudan a la identificación y, por otro lado, los modelos del resto de bases de datos han sido entrenados con un contenido *fake* más elaborado, ya que han sido preparados para escenarios más difíciles de detectar. Por otra parte, la base de datos ***FaceForensics++*** corresponde con la más difícil de detectar para el resto de modelos. La principal razón reside en cómo se han generado estos vídeos *fake*: en este caso, la técnica utilizada ha sido *Face Swap*, en vez de *DeepFakes*, como en el resto de bases de datos. Por último, los modelos de la 2ª generación ofrecen rendimientos por encima del **70 %** de AUC (algunos incluso

superiores al **80 %**) al evaluar las bases de datos de esa misma generación. Los modelos de *DFDC* consiguen mejores resultados al evaluar la base de datos *Celeb-DF* que al revés -los modelos de *Celeb-DF* al evaluar la base de datos *DFDC*-, lo que indica lo comentado en la anterior sección: ***DFDC*** es la la base de datos más **difícil** de detectar hasta el momento.

5

Conclusiones Finales y Trabajo Futuro

5.1. Conclusiones

El estudio realizado en el presente Trabajo de Fin de Grado se ha centrado en el análisis a nivel de rendimiento e imagen de la técnicas de manipulación digital englobadas dentro de *Identity Swap*. Para ello, se ha diseñado una arquitectura con dos módulos: *i)* un sistema basado en la segmentación de regiones faciales a partir de los vídeos procedentes de las bases de datos más destacadas, y *ii)* tres sistemas basados en la detección *fake* de las diferentes zonas faciales obtenidas.

En primer lugar, este trabajo se ha enfocado en el **análisis de las diferentes regiones de la cara** para observar el rendimiento de cada una de ellas. En efecto, el estudio ha mostrado la evolución de las técnicas de manipulación a través de las bases de datos públicas, donde se ha percibido una mejora en las distintas regiones faciales analizadas.

En segundo lugar, para el entrenamiento y posterior evaluación de los distintos modelos faciales se han implementado tres **sistemas de detección *fake* en el estado del arte**: *XceptionNet*, *Capsule Forensics* y DSP-FWA. La arquitectura de cada uno de estos sistemas ha sido un factor clave a la hora de analizar los rendimientos conseguidos, ya que cada uno ha aportado una perspectiva diferente para cada base de datos, aunque mayoritariamente se hayan obtenido valores muy similares entre sistemas.

Como resultado del estudio, se ha observado una mejora notable entre las bases de datos analizadas en este trabajo: los sistemas de detección han obtenido unos rendimientos superiores en comparación con las bases de datos de la 1ª generación -formadas por *UADFV* y *FaceForensics++*- y la 2ª generación -formadas por *Celeb-DF* y *DFDC*-, lo que ha supuesto un progreso en las técnicas de manipulación basadas en *Identity Swap*. Además, debido al estudio basado en modelos de regiones faciales, se ha podido analizar la evolución que ha tenido cada una de ellas. Por un lado, regiones como la nariz o el resto de la cara han obtenido unos rendimientos muy inferiores, lo que indica una mejora en la manipulación de estas zonas. Por otro lado, los sistemas han logrado alcanzar mejores rendimientos en la zona de los ojos, lo que señala que esta zona es más difícil de manipular y por tanto, supone un desafío para las próximas técnicas de manipulación.

Finalmente, estos resultados han generado un artículo que se ha publicado al *IEEE Interna-*

tional Conference on Pattern Recognition 2020.

5.2. Trabajo Futuro

A lo largo del estudio realizado en este proyecto, han ido surgiendo nuevas líneas de investigación que pueden dar lugar a trabajos futuros:

- Implementación de una arquitectura basada en Redes Neuronales Recurrentes (RNN). En este estudio, el análisis se ha centrado a nivel de imagen con CNN, sin embargo, el rendimiento de estos sistemas pueden mejorar si se considera la información temporal entre *frames* del vídeo usando RNN.
- Análisis a nivel de rendimiento de las técnicas *DeepFakes* en función del sexo y etnia de una persona, ya que los estudios preliminares indican que existen diferencias importantes.
- Implementación de un sistema de detección *fake* basado en la fusión de varias regiones faciales -hasta ahora, el estudio se ha centrado en función del tipo de técnica utilizada durante la manipulación así como la calidad y escenarios del vídeo-.
- Estudio en profundidad de los sistemas *Capsule Forensics* y *DSP-FWA* con el objetivo de implementar mejoras.

Glosario de Acrónimos

- **GAN**: *Generative Adversarial Network*
- **CNN**: *Convolutional Neural Network*
- **DFDC**: *DeepFake Detection Challenge*
- **FF++**: *FaceForensics++*
- **Grad-CAM**: *Gradient-weighted Class Activation Mapping*
- **AUC**: *Area Under the Curve*
- **SVM**: *Support Vector Machine*
- **VPR**: Razón de Verdaderos Positivos
- **FPR**: Razón de Falsos Positivos
- **FA**: *False Acceptance*
- **FR**: *False Rejection*
- **EER**: *Equal Error Rate*
- **DSP**: *Dual Spatial Pyramid*
- **FWA**: *Face Warping Artifacts*
- **SPP**: *Spatial Pyramid Pooling*
- **RNN**: *Recurrent Neural Network*

Bibliografía

[1] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild et al. The Science of Fake News. *Science*, 359(6380):1094–1096, 2018.

[2] R. Tolosana , R. Vera-Rodriguez, J. Fiérrez, A. Morales and J. Ortega-García. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *arXiv preprint arXiv:2001.00179*, abs/2001.00179, 2020.

[3] J. Stehouwer, H. Dang, F. Liu, X. Liu, A. Jain. On the Detection of Digital Face Manipulation. *arXiv preprint arXiv:1910.01717*, 2019.

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proc. International Conference on Computer Vision*, 2019.

[5] DB. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Ferrer. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[6] K. Schoolov. How Facebook, Twitter and Google are Working to Prevent Deepfakes from Fooling You, 2019. Available: https://www.cnbc.com/2019/09/29/how-facebook-twitter-and-google-work-to-detect-and-prevent-deepfakes.html.

[7] A. Hutchinson. Snapchat and TikTok are Both Reportedly Working on New 'Deepfake' Type Features, 2020. Available: https://www.socialmediatoday.com/news/snapchat-and-tiktok-are-both-reportedly-working-on-new-deepfake-type-feat/569792/.

[8] P. Fraga-Lamas and T. M. Fernandez-Carames. Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality. *IT Professional*, 22(2):53–59, 2020.

[9] R. Tolosana, S. Romero-Tapiador, J. Fierrez and R. Vera-Rodriguez. DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. *arXiv preprint arXiv: 2004.07532*, 2020.

[10] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L.P. Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proc. International Conference on Automatic Face Gesture Recognition*. IEEE, 2018.

[11] I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[12] D. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative Adversarial Nets. In *Proc. Advances in Neural Information Processing Systems*, 2014.

[14] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do GANs Leave Artificial Fingerprints? In *Proc. IEEE Conference on Multime- dia Information Processing and Retrieval*, 2019.

[15] J. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença and J. Fierrez. GANprintR: Improved Fakes and Evaluation of the State-of-the-Art in Face Manipulation Detection. *arXiv preprint arXiv:1911.05351*, 2019.

[16] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[17] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[18] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.

[19] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.

[20] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Deferred Neural Rendering: Image Synthesis using Neural Textures. *ACM Transactions on Graphics*, 38(4):1–12, 2019.

[21] G. Wolberg. Image Morphing: A Survey. *The Visual Computer*, 14, 1999.

[22] R. Gross and L. Sweeney and F. de la Torre and S. Baker. Model-Based Face De-Identification. In *Proc. Conference on Computer Vision and Pattern Recognition Workshop*, 2006.

[23] S. Agarwal, H. Farid, O. Fried, and M. Agrawala. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[24] P. Korshunov and S. Marcel. Deepfakes: a New Threat to Face Recognition? Assessment and Detection, 2018.

[25] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proc. International Conference on Computer Vision*, 2017.

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499, 2016.

[27] Y. Li, M. Chang, and S. Lyu. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *Proc. International Workshop on Information Forensics and Security*, 2018.

[28] Google AI. Contributing Data to Deepfake Detection Research, 2019. https://ai.googleblog.com/2019/ 09/contributing-data-to-deepfake-detection.html.

[29] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-DF: A New Dataset for DeepFake Forensics, 2019.

[30] C. Sanderson and B. Lovell. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. In *Proc. International Conference on Biometrics*, 2009.

[31] S. Narkhede. Understanding AUC - ROC Curve, 2018. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[32] C. González. Adaptación de Sistemas de Verificación de Firma Manuscrita a Dispositivos Móviles. *TFG. Escuela Politécnica Superior, Universidad Autónoma de Madrid*, 2019.

[33] F. Matern, C. Riess, and M. Stamminger. Exploiting Visual Artifacts to Expose DeepFakes and Face Manipulations. In *Proc. Winter Applications of Computer Vision Workshops*, 2019.

[34] X. Yang, Y. Li, and S. Lyu. Exposing Deep Fakes Using Inconsistent Head Poses. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2019.

[35] Y. Li and S. Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[36] H.H. Nguyen, J. Yamagishi and I. Echizen. Use of a Capsule Network to Detect Fake Images and Videos. *arXiv preprint arXiv:1910.12467*, 2019.

[37] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[38] J. Galbally and S. Marcel and J. Fierrez. Image Quality Assessment for Fake Biometric Detection: Application to Iris, Fingerprint, and Face Recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, 2014.

[39] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2015.

[41] K. P. Mensah, F. A. Adebayo, A. M. Ayidzoe and E. Baagyire Y. Capsule Networks – A Survey. *Journal of King Saud University - Computer and Information Sciences*, 2019.

[42] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[43] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2009.

[44] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.

[45] K. He, X. Zhang, S. Ren, and J. Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[46] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proc. International Conference on Computer Vision*, 2017.